Form of paper: Regular Paper

Field and Topic: Biochemistry, Protein Structure, Enzyme Inhibitors

Structure models of G72, the product of a susceptibility gene to schizophrenia

Yusuke Kato and Kiyoshi Fukui*

Affiliation

Division of Enzyme Pathophysiology, Institute for Enzyme Research, Tokushima University, Tokushima 770-8503, Japan

*Correspondence to:

Kiyoshi Fukui, Division of Enzyme Pathophysiology, Institute for Enzyme Research, Tokushima University, 3-18-15 Kuramoto, Tokushima 770-8503, Japan. Tel: +81-88-633-7430, Fax: +81-88-633-7431, E-mail: kiyo.fukui@tokushima-u.ac.jp

Running Title: Structure models of G72

Abbreviations:

CD	C-terminal domain
DAO	D-amino acid oxidase
Gαq	α subunit of guanine nucleotide-binding protein, Gq
GRK	G-protein coupled receptor kinases
HsdM	modification subunit of Type I DNA methyltransferases
MD	Molecular dynamics
ND	N-terminal domain
NMDA	N-methyl-D-aspartate
OGT	*O*-GlcNAc transferases
RH	regulator of G protein signaling homology
RMSD	root-mean-square deviation

Summary

The G72 gene is one of the most susceptible genes to schizophrenia and is contained exclusively in the genomes of primates. The product of the G72 gene modulates the activity of D-amino acid oxidase (DAO) and is a small protein prone to aggregate, which hampers its structural studies. In addition, lack of a known structure of a homologue makes it difficult to use the homology modeling method for the prediction of the structure. Thus, we first developed a hybrid *ab initio* approach for small proteins prior to the prediction of the structure of G72. The approach uses three known *ab initio* algorithms. To evaluate the hybrid approach, we tested our prediction of the structure of the amino acid sequences whose structures were already solved and compared the predicted structures with the experimentally solved structures. Based on these comparisons, the average accuracy of our approach was calculated to be ~5 Å. We then applied the approach to the sequence of G72 and successfully predicted the structures of the N- and C-terminal domains (ND and CD, respectively) of G72. The predicted structures of ND and CD were similar to membrane-bound proteins and adaptor proteins, respectively.

Key words:

G72/*ab initio* method/structure prediction/schizophrenia/D-amino acid oxidase

Introduction

About 1% of world population develops schizophrenia (*1*). Among all schizophrenia linkage regions, SCZD7 on chromosome 13q32-q33 (MIM 603176) is one of the most important regions (*1, 2*). Chumakov *et al.* reported overlapping genes including G30 and G72 in this region and an associated common SNP that replaces Arg30 of the G72 protein with Lys (*3*). In addition, quite small p value for the association of polymorphism in the G30/G72 locus (e.g. rs4517638 $p < 0.00002$) was reported according to recent GWAS studies (*4*). Transgenic mice of G72 showed behavioral alterations indicative of psychiatric disorders including abnormal motor coordination phenotype and deficits in prepulse inhibition (*5*). Moreover, prepulse inhibition deficits were normalized by the administration of haloperidol, an antagonist of the dopamine receptor D_2. These findings indicate that the G30/G72 locus is the true-positive and robust region associated to schizophrenia. Furthermore, it has been suggested that the G30/G72 locus is associated with other psychiatric disorders including bipolar disorder and Alzheimer's disease (*6, 7*).

The G72 protein is primate-specific and has several variants. The 153 residue variant has been found exclusively in human. The amino acid sequence of G72 lacks recognizable protein motifs that are common to other proteins. The G72 protein interacts with D-amino acid oxidase (DAO) and regulates the activity of DAO (*3, 8, 9*). DAO regulates the amount of intracerebral D-Ser, one of the major co-agonists of the N-methyl-D-aspartate (NMDA) receptors (*10*). Thus, dysfunction of G72 may cause the hypofunction of the NMDA receptors, leading to the onset of schizophrenia. This is in line with the glutamate hypothesis of schizophrenia (*11*).

G72 also regulates the function of mitochondria by promoting mitochondrial fragmentation and dendritic arborization (*12*). Changes in the redox states were suggested based on various evidences including decrease in the glutathione levels in brains of patients and transgenic mice of G72 and improved mismatch negativity of patients after the treatment with reducing agents (*13-15*). These suggest that aberration of the redox state and mitochondria due to the dysfunction of G72 may cause schizophrenia.

Because the structures of the G72 protein and its homologues have been unknown, the molecular mechanism and function of G72 are still elusive. It is difficult to apply homology modeling to predict the structure of G72 without that of a homologue. When we have no structure of a homologue, the *ab initio* methods may be useful to predict protein structures. Although the accuracy of the *ab initio* methods has been improved rapidly (*16-19*), most of known *ab initio* methods still do not predict the one best structure but do predict multiple possible structures for a query amino acid sequence. To solve this problem, we first developed a hybrid *ab initio* approach to predict the structures of small helical basic proteins and then applied it to predict the structure of G72.

Materials and methods

Secondary Structure prediction

We predicted the secondary structure of G72 with the JPRED and PSIPRED algorithms prior to the prediction of the tertiary structure (*20, 21*).

Structure prediction with the hybrid ab initio approach

We produced 5 candidate models with I-TASSER, 10 candidate models with QUARK and ~20000 decoy structures with the AbinitoRelax algorithm of the Rosetta suite using the sequences of the PDB files of 2QFF, 3FEA, 2EFV, 1MN8, 1O82 and 2ZKO (*16, 18, 19*). We chose the six sequences from the proteins that were used for the evaluation of the QUARK program based on the following criteria: helical proteins with less than 100 residues, positive isoelectric points and no prosthetic group (*18*). No structural information of these PDB files and their homologues was used for the prediction. Two rounds of clustering using the cluster algorithm of the Rosetta suite were performed to obtain the consensus clusters that contained the models and decoys from all of the three algorithms. For the final selection of the best model, the models and decoys in the consensus clusters were directly compared by calculating the RMSD values with the Swiss PDB Viewer by excluding 10 residues from N- and C-termini (*22*).

Model building of ND and CD of G72 was performed with the same approach using the regions #1 – 71 and #72 – 153 of the 153 residue-long isoform. Prior to the model building, the sequence was analyzed with NCBI BLAST to define the domains based on Conserved Domain Database (*23*).

Molecular dynamics

The best ND and CD models were subject to MD simulations. The AMBER ff14SB force field was applied (*24*). The domains were placed in a periodic TIP3P box with the LEaP module of AMBER 14 (*25*). The systems were neutralized by replacing TIP3P models with Cl^- ions. A 10 Å cut-off was applied to the non-bonded interactions according to the Lennard–Jones potential. Steepest descent minimization was performed followed by conjugate gradient minimization with the Particle Mesh Ewald method with constant-volume periodic boundaries and with position restraints for the protein atoms. A 200 ps heating procedure was performed from 0 to 300 K under constant volume periodic boundaries. The step size was set to 0.002 ps. Equilibration and production MD was performed with a constant pressure periodic boundary at 1 atm and 300 K without position restraints. The duration of the production MD was 10 ns and 30 ns for ND and CD, respectively. Evaluation of the model structures was performed with the Ramachandran plot, Verify3D and ERRAT (*26-28*).

Results

The hybrid ab initio approach

Each *ab initio* prediction algorithm such as I-TASSER, QUARK and AbinitoRelax of the Rosetta suite produced multiple candidate models and decoys for an amino acid sequence without template structures of homologues (**Fig. 1A**). We classified all these models/decoys with two rounds of clustering and found consensus clusters that contained models/decoys from all the three algorithms. The models/decoys in the consensus clusters were compared with each other to determine the final model. We call this approach the hybrid *ab initio* approach. We examined the accuracy of our approach using six sequences of deposited structures in Protein Data Bank (**Fig. 1B**). To this end, we built the structural models of these sequences without using structural information of these proteins and their homologues. In the present study, we focused on the prediction of structures of helical proteins with less than 100 residues, positive isoelectric points and no prosthetic group. All of the six proteins for the test satisfy the criteria. We focused on the helical proteins because the secondary structure prediction suggested that G72 is helix-rich (supplementary **Fig. S1**). After the selection of the final models, we compared the predicted structures and those deposited in Protein Data Bank. The average of the backbone root-mean-square deviation (RMSD) values between the predicted and deposited structures was ~5 Å, which is close to the size of a small amino acid. Moreover, the average RMSD value further improved when 10 residues from N- and C-termini were excluded from the calculation of the RMSD of each comparison. These suggested that the core regions of the predicted structures were predicted more correctly than the terminal regions and that our approach successfully predicted correct folds (**Fig. 1B, C**). It is notable that our approach succeeded in choosing the one best model for each amino acid sequence out of multiple candidate models/decoys.

Domain composition of the G72 protein

Three variants of human G72 transcripts have been reported. Expression of the 153 residue-long variant was observed in the human brain cortex (*8*). Only the 153 residue-long variant has been reported to interact with DAO (*3, 9*). In addition, the reported amino acid substitution, Arg30Lys, associated with the disorder is that of the 153 residue variant. Conserved Domain Database (*23*), which is a resource of the National Center for Biotechnology Information (NCBI), identified a domain from the residue number 72 to 153 within the 153 residue variant (**Fig. 2**). We call this region the C-terminal domain (CD) and call the remaining region the N-terminal domain (ND). Most of the sequence of CD is conserved among the three human variants, whereas that of ND is not.

Modeling of ND

Structure prediction of ND was performed with the hybrid *ab initio* approach. We clustered models/decoys of ND after building models/decoys. 11 clusters were found in the first round of clustering (**Fig. 3A**). Two of them were consensus clusters in which the models/decoys produced by the three algorithms were included. After the second round of clustering for the consensus clusters, we found three consensus clusters. We compared the decoys/models from the three different algorithms within each of the consensus clusters by calculating the backbone RMSD between the decoys/models. Subsequently, we calculated the average of the RMSD values for each decoy/model as shown in **Fig. 3C** to choose the decoy/model with the smallest average RMSD as the best model. The best model was the 4th model of the 10 models that were originally produced by QUARK (Q4). We confirmed that the Q4 model shared the similar overall structure with models/decoys produced by AbinitoRelax (AR c.0.3) and I-TASSER (IT1) (**Fig. 3B, C**). The best model was subject to molecular dynamics (MD) simulation and reached equilibrium after ~0.5 ns, suggesting that this model structure is stable in aqueous solution (**Supplementary Fig. S2)**. Evaluation by the Ramachandran plot, Verify3D and ERRAT indicated acceptable profiles (**Table 1**). The RMSD value between Q4 and IT1 was 3.75 Å, whereas the value between Q4 and AR c.0.3 was 3.93 Å. These indicate diversity of the model structures produced by the different algorithms. Such an extent of the diversity is acceptable because these RMSD values were comparable to the average accuracy of the hybrid *ab initio* approach (~ 5 Å). The RMSD values in Supplementary **Fig. S2** indicate the extent of conformational change in the course of the MD simulation with Q4 as the initial structure and were roughly 1 to 2 Å after ~0.5 ns. It is possible that the equilibrated conformations after ~0.5 ns may be those that were trapped in local minima of the energy landscape. However, it is predicted that the RMSD of these conformations with respect to the true structure should be less than ~ 5 Å based on the accuracy of the hybrid *ab initio* approach.

Modeling of CD

Modeling of CD was performed similarly to that of ND. The consensus clusters contained the models/decoys from all the three algorithms. The final comparison of the best models from the three algorithms indicated that the 8th model produced by QUARK (Q8) should be chosen as the final model (**Fig. 4**). This model shared the overall structure with the models/decoys from AbinitoRelax and I-TASSER (AR c.0.4 and IT2, respectively). The best model was subject to MD simulation and reached equilibrium after ~5 ns (**Supplementary Fig. S3)**, suggesting that this model structure is stable. Evaluation by the Ramachandran plot, Verify3D and ERRAT indicated acceptable profiles (**Table 2**). The RMSD value between Q8 and IT2 was 4.57 Å, whereas the value between Q8 and AR c.0.4 was 4.69 Å. The RMSD values in Supplementary **Fig. S3** indicate the extent of conformational change in the course of the MD simulation with Q8 as the initial structure and were roughly 3 to 4 Å after ~5 ns. Although it is possible that these conformations may be trapped in local minima of the energy landscape, it is predicted that the RMSD with respect to the true structure should be less than ~5 Å as was the case of ND.

Distribution of surface charges

Both ND and CD showed different surface charge distribution (**Fig. 5**). Notably, CD contains clusters of opposite charges on opposite sides of the molecule. The 30th residue is located at the center of the positively charged cluster of ND. Arg30 is on the same face as the positive charges including those of Lys4, Lys36, Arg57, Arg64 and the N-terminus (**Fig. 5A**).

Fold search

We performed a fold search with COFACTOR using the best ND model as a query structure and obtained the PDB codes of 10 structural analogues (**Supplementary Table S1**). Direct observation of these analogues confirmed that the analogs from Rank 1 to 4 appeared similar to ND (**Fig. 6A, B**), whereas the others contained inconsistent paths of polypeptide chains compared with ND (**Fig. 6C**). The proteins that ranked number 1 and 3 (PDB code: 4X82 and 5AEZ, respectively) were membrane transporters. 4X82 is the PDB code of the extracellular domain of a Zn transporter, whereas 5AEZ is that of the transmembrane domain of an ammonium transporter. In addition, we obtained 110 similar structures to ND from a fold search using the Dali server (**Supplementary Table S2**). We confirmed that top ~80 structures appeared similar to ND by direct observation. Most of the ~80 structures were those of *O*-GlcNAc transferases (OGT). The part of OGT that resembles ND binds to the phosphate groups of UDP.

A search for 10 structural analogues of the CD model with COFACTOR was performed (**Supplementary Table S3**). We confirmed that the analogues that ranked top, fifth and sixth had similar structures to CD by direct observation of the superposition of the structures of the CD model and the hit proteins, but the other analogues did not (**Figs. 7A-E**). The top-ranking PDB code 2OKC is that of the modification subunit (HsdM) of Type I DNA methyltransferases. A fold search with the Dali server was also preformed using the CD model. 26 similar structures were obtained from the search (**Supplementary Table S4**). 18 of the 26 structures were those of regulator of G protein signaling homology bundle subdomain (RH bundle subdomain) of G-protein coupled receptor kinases (GRK). We confirmed that the RH bundle subdomains are similar to CD by observing the structures of those proteins (**Fig. 7D, F**).

Discussion

We developed the hybrid *ab initio* approach for the positively charged small proteins that is rich in α-helices. The average backbone accuracy of the approach approximately corresponds to the size of a small amino acid. Moreover, our approach succeeded in choosing the one best structural model for an amino acid sequence out of multiple decoys/models that were produced by the known algorithms. We predicted the structures of ND and CD of G72 using the approach. The predicted structures of ND and CD were rich in α-helices, which was consistent with the secondary structure contents data that were analyzed with circular dichroism (*29*). The surface charges of ND and CD were rich in positive charges, which may be advantageous to the interaction with human DAO whose surface is negatively charged (*30, 31*). Arg30Lys is the susceptibility substitution of the G72 protein. It is therefore speculated that the substitution may have an impact on the interaction with DAO because the arginine is at the center of the positive charge cluster. It is intriguing that the predicted structure of ND is similar to parts of membrane proteins and OGTs. These might suggest that ND can locate on the surface of membranes of organelles including mitochondria. Indeed, G72 was reported to locate on mitochondria and in cytosol (*8, 12*). We speculate that ND might bind to the phosphate groups of membrane lipids, because the part of OGT that resembles ND binds to the phosphate groups of UDP. ND is too small to be a transmembrane protein as shown in **Fig 6B**. In addition, the hydropathy plot did not indicate a transmembrane region in the G72 sequence.

The fold searches of CD indicated that CD is similar to a part of HsdM and the RH bundle subdomain. The HsdM molecules form a homodimer (*32*). The predicted structure of CD is similar to the helix bundle that serves as the dimer interface of HsdM. The RH bundle subdomain functions as the binding interface with the α subunit of guanine nucleotide-binding protein, Gq (Gαq), to inhibit G-proteins (*33*). It was reported that two G72 molecules and four DAO molecules form a complex (*8*). Thus, CD might be an interface for complex formation with the other proteins including DAO, which might suggest that G72 serves as an adaptor protein.

The present study developed a hybrid approach to predict protein structures and predicted the structures of the domains of the G72 protein. This approach may be useful for the prediction of structures of small nucleic acid-binding proteins, because many of these proteins are helical and positively charged. The application of this approach might be expanded to negatively charged and/or non-helical proteins in the future. The predicted structures may be useful for the functional analysis of G72 and as the target structures for the development of psychopharmaceutical drugs because mutations in G72 are presumed to contribute to the development of psychiatric disorders including schizophrenia and bipolar disorder. However, it is important to improve the accuracy of the predicted structures for efficient drug screening. The MD simulation with microsecond- to millisecond-scale may improve the accuracy of the models. The other possibility to improve the accuracy is to combine experimental data into the prediction calculation. By identifying cross-linked residues with Liquid Chromatography Mass Spectrometry after chemical cross-linking, it is possible to obtain distance restraints for the calculation.

Supplementary Data

Supplementary Data are available at JB Online.

Funding

This work was supported in part by a grant for Enzyme Research from the Japan Foundation for Applied Enzymology. We thank Ms. Maki Kato and the other contributors for crowdfunding the present research launched by Academist Inc (Tokyo, Japan).

Conflict of Interest

None declared.

Acknowledgments

We thank the members of Division of Enzyme Pathophysiology (Tokushima University) for helpful discussion. Authors declare no conflict of interests.

References

1. Drews, E., Otte, D.M., and Zimmer, A. (2013) Involvement of the primate specific gene G72 in schizophrenia: From genetic studies to pathomechanisms. Neurosci Biobehav Rev 37, 2410-2417

2. Abou Jamra, R., Schmael, C., Cichon, S., Rietschel, M., Schumacher, J., and Nothen, M.M. (2006) The G72/G30 gene locus in psychiatric disorders: a challenge to diagnostic

boundaries? Schizophr Bull 32, 599-608

3. Chumakov, I., Blumenfeld, M., Guerassimenko, O., Cavarec, L., Palicio, M., Abderrahim, H., Bougueleret, L., Barry, C., Tanaka, H., La Rosa, P., Puech, A., Tahri, N., Cohen-Akenine, A., Delabrosse, S., Lissarrague, S., Picard, F.P., Maurice, K., Essioux, L., Millasseau, P., Grel, P., Debailleul, V., Simon, A.M., Caterina, D., Dufaure, I., Malekzadeh, K., Belova, M., Luan, J.J., Bouillot, M., Sambucy, J.L., Primas, G., Saumier, M., Boubkiri, N., Martin-Saumier, S., Nasroune, M., Peixoto, H., Delaye, A., Pinchot, V., Bastucci, M., Guillou, S., Chevillon, M., Sainz-Fuertes, R., Meguenni, S., Aurich-Costa, J., Cherif, D., Gimalac, A., Van Duijn, C., Gauvreau, D., Ouellette, G., Fortier, I., Raelson, J., Sherbatich, T., Riazanskaia, N., Rogaev, E., Raeymaekers, P., Aerssens, J., Konings, F., Luyten, W., Macciardi, F., Sham, P.C., Straub, R.E., Weinberger, D.R., Cohen, N., and Cohen, D. (2002) Genetic and physiological data implicating the new human gene G72 and the gene for D-amino acid oxidase in schizophrenia. Proc Natl Acad Sci U S A 99, 13675-13680

4. Ripke S, S.A., Kendler KS, Levinson DF, Sklar P, Holmans PA, Lin DY, Duan J, Ophoff RA, Andreassen OA, Scolnick E, Cichon S, St Clair D, Corvin A, Gurling H, Werge T, Rujescu D, Blackwood DH, Pato CN, Malhotra AK, Purcell S, Dudbridge F, Neale BM, Rossin L, Visscher PM, Posthuma D, Ruderfer DM, Fanous A, Stefansson H, Steinberg S, Mowry BJ, Golimbet V, De Hert M, Jönsson EG, Bitter I, Pietiläinen OP, Collier DA, Tosato S, Agartz I, Albus M, Alexander M, Amdur RL, Amin F, Bass N, Bergen SE, Black DW, Børglum AD, Brown MA, Bruggeman R, Buccola NG, Byerley WF, Cahn W, Cantor RM, Carr VJ, Catts SV, Choudhury K, Cloninger CR, Cormican P, Craddock N, Danoy PA, Datta S, de Hann L, Demontis D, Dikeos D, Djurovic S, Donnelly P, Donohoe G, Duong L, Dwyer S, Fink-Jensen A, Freedman R, Freimer NB, Friedl M, Georgieva L, Giegling I, Gill M, Glenthøj B, Godard S, Hamshere M, Hansen M, Hansen T, Hartmann AM, Henskens FA, Hougaard DM, Hultman CM, Ingason A, Jablensky AV, Jakobsen KD, Jay M, Jürgens G, Kahn RS, Keller MC, Kenis G, Kenny E, Kim Y, Kirov GK, Konnerth H, Konte B, Krabbendam L, Krausucki R, Lasseter VK, Laurent C, Lawrence J, Lencz T, Lerer FB, Liang KY, Lichtenstein P, Lieberman JA, Linszen DH, Lönnqvist J, Loughland CM, Maclean AW, Maher BS, Maier W, Mallet J, Malloy P, Mattheisen M, Mattinsgsdal M, McGhee KA, McGrath JJ, McIntosh A, McLean DE, McQuillin A, Melle I, Michie PT, Milanova V, Morris DW, Mors O, Mortensen PB, Moskvina V, Muglia P, Myin-Germeys I, Nertney DA, Nestadt G, Nielsen J, Nikolov I, Nordentroft M, Norton N, Nöthen MM, O'Dushlaine CT, Olincy A, Olsen L, O'Neill FA, Ørntoft T, Owen MJ, Pantelis C,

Papadimitriou G, Pato MT, Peltonen L, Petursson H, Pickard B, Pimm J, Pulver AE, Puri V, Quested D, Quinn EM, Rasmussen HB, Réthelyi JM, Ribble R, Rietschel M, Riley BP, Ruggeri M, Schall U, Schulze TG, Schwab SG, Scott RJ, Shi J, Sigurdsson E, Silverman JM, Spencer CC, Stefansson K, Strange A, Strengman E, Stroup TS, Suvisaari J, Tereniuis L, Thirumalai S, Thygesen JH, Timm S, Toncheva D, van den Oord E, van Os J, van Winkel R, Veldink J, Walsh D, Wang AG, Wiersma D, Wildenauer DB, Williams HJ, Williams NM, Wormley B, Zammit S, Sullivan PF, O'Donovan MC, Daly MJ, Gejman PV. (2011) Genome-wide association study identifies five new schizophrenia loci. Nat Genet 43, 969-976

5. Otte, D.M., Bilkei-Gorzo, A., Filiou, M.D., Turck, C.W., Yilmaz, O., Holst, M.I., Schilling, K., Abou-Jamra, R., Schumacher, J., Benzel, I., Kunz, W.S., Beck, H., and Zimmer, A. (2009) Behavioral changes in G72/G30 transgenic mice. Eur Neuropsychopharmacol 19, 339-348
6. Liu, C., Badner, J.A., Christian, S.L., Guroff, J.J., Detera-Wadleigh, S.D., and Gershon, E.S. (2001) Fine mapping supports previous linkage evidence for a bipolar disorder susceptibility locus on 13q32. Am J Med Genet 105, 375-380
7. Velez, J.I., Rivera, D., Mastronardi, C.A., Patel, H.R., Tobon, C., Villegas, A., Cai, Y., Easteal, S., Lopera, F., and Arcos-Burgos, M. (2016) A Mutation in DAOA Modifies the Age of Onset in PSEN1 E280A Alzheimer's Disease. Neural Plast 2016, 9760314
8. Sacchi, S., Bernasconi, M., Martineau, M., Mothet, J.P., Ruzzene, M., Pilone, M.S., Pollegioni, L., and Molla, G. (2008) pLG72 modulates intracellular D-serine levels through its interaction with D-amino acid oxidase: effect on schizophrenia susceptibility. J Biol Chem 283, 22244-22256
9. Chang, S.L., Hsieh, C.H., Chen, Y.J., Wang, C.M., Shih, C.S., Huang, P.W., Mir, A., Lane, H.Y., Tsai, G.E., and Chang, H.T. (2014) The C-terminal region of G72 increases D-amino acid oxidase activity. Int J Mol Sci 15, 29-43
10. Morikawa, A., Hamase, K., Inoue, T., Konno, R., Niwa, A., and Zaitsu, K. (2001) Determination of free D-aspartic acid, D-serine and D-alanine in the brain of mutant mice lacking D-amino acid oxidase activity. J Chromatogr B Biomed Sci Appl 757, 119-125
11. Lisman, J.E., Coyle, J.T., Green, R.W., Javitt, D.C., Benes, F.M., Heckers, S., and Grace, A.A. (2008) Circuit-based framework for understanding neurotransmitter and risk gene interactions in schizophrenia. Trends Neurosci 31, 234-242
12. Kvajo, M., Dhilla, A., Swor, D.E., Karayiorgou, M., and Gogos, J.A. (2008) Evidence

implicating the candidate schizophrenia/bipolar disorder susceptibility gene G72 in mitochondrial function. Mol Psychiatry 13, 685-696

13. Do, K.Q., Trabesinger, A.H., Kirsten-Kruger, M., Lauer, C.J., Dydak, U., Hell, D., Holsboer, F., Boesiger, P., and Cuenod, M. (2000) Schizophrenia: glutathione deficit in cerebrospinal fluid and prefrontal cortex *in vivo*. Eur J Neurosci 12, 3721-3728

14. Lavoie, S., Murray, M.M., Deppen, P., Knyazeva, M.G., Berk, M., Boulat, O., Bovet, P., Bush, A.I., Conus, P., Copolov, D., Fornari, E., Meuli, R., Solida, A., Vianin, P., Cuenod, M., Buclin, T., and Do, K.Q. (2008) Glutathione precursor, N-acetyl-cysteine, improves mismatch negativity in schizophrenia patients. Neuropsychopharmacology 33, 2187-2199

15. Otte, D.M., Sommersberg, B., Kudin, A., Guerrero, C., Albayram, O., Filiou, M.D., Frisch, P., Yilmaz, O., Drews, E., Turck, C.W., Bilkei-Gorzo, A., Kunz, W.S., Beck, H., and Zimmer, A. (2011) N-acetyl cysteine treatment rescues cognitive deficits induced by mitochondrial dysfunction in G72/G30 transgenic mice. Neuropsychopharmacology 36, 2233-2243

16. Bradley, P., Misura, K.M., and Baker, D. (2005) Toward high-resolution de novo structure prediction for small proteins. Science 309, 1868-1871

17. Kim, H., and Kihara, D. (2015) Protein structure prediction using residue- and fragment-environment potentials in CASP11. Proteins 2015

18. Xu, D., and Zhang, Y. (2012) Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. Proteins 80, 1715-1735

19. Roy, A., Kucukural, A., and Zhang, Y. (2010) I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc 5, 725-738

20. Drozdetskiy, A., Cole, C., Procter, J., and Barton, G.J. (2015) JPred4: a protein secondary structure prediction server. Nucleic acids res 43, W389-394

21. McGuffin, L.J., Bryson, K., and Jones, D.T. (2000) The PSIPRED protein structure prediction server. Bioinformatics 16, 404-405

22. Guex, N., and Peitsch, M.C. (1997) SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 18, 2714-2723

23. Marchler-Bauer, A., Anderson, J.B., Chitsaz, F., Derbyshire, M.K., DeWeese-Scott, C., Fong, J.H., Geer, L.Y., Geer, R.C., Gonzales, N.R., Gwadz, M., He, S., Hurwitz, D.I., Jackson, J.D., Ke, Z., Lanczycki, C.J., Liebert, C.A., Liu, C., Lu, F., Lu, S., Marchler, G.H., Mullokandov, M., Song, J.S., Tasneem, A., Thanki, N., Yamashita, R.A., Zhang, D., Zhang, N., and Bryant, S.H. (2009) CDD: specific functional annotation with the Conserved

Domain Database. Nucleic Acids Res 37, D205-210

24. Maier, J.A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K.E., and Simmerling, C. (2015) ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. J Chem Theory Comput 11, 3696-3713

25. Case, D.A., Babin, V., J.T. Berryman, R.M. Betz, Cai, Q., Cerutti, D.S., Cheatham, I., T.E. , Darden, T.A., Duke, R.E., Gohlke, H., Goetz, A.W., Gusarov, S., Homeyer, N., Janowski, P., Kaus, J., Kolossvary, I., Kovalenko, A., Lee, T.S., LeGrand, S., Luchko, T., Luo, R., Madej, B., Merz, K.M., Paesani, F., Roe, D.R., Roitberg, A., Sagui, C., Salomon-Ferrer, R., Seabra, G., Simmerling, C.L., Smith, W., Swails, J., Walker, R.C., Wang, J., Wolf, R.M., Wu, X., and Kollman, P.A. (2014) Amber 14, University of California, San Francisco,

26. Lovell, S.C., Davis, I.W., Arendall, W.B., 3rd, de Bakker, P.I., Word, J.M., Prisant, M.G., Richardson, J.S., and Richardson, D.C. (2003) Structure validation by Calpha geometry: phi,psi and Cbeta deviation. Proteins 50, 437-450

27. Luthy, R., Bowie, J.U., and Eisenberg, D. (1992) Assessment of protein models with three-dimensional profiles. Nature 356, 83-85

28. Colovos, C., and Yeates, T.O. (1993) Verification of protein structures: patterns of nonbonded atomic interactions. Protein Sci 2, 1511-1519

29. Molla, G., Bernasconi, M., Sacchi, S., Pilone, M.S., and Pollegioni, L. (2006) Expression in Escherichia coli and *in vitro* refolding of the human protein pLG72. Protein Expr Purif 46, 150-155

30. Kawazoe, T., Tsuge, H., Pilone, M.S., and Fukui, K. (2006) Crystal structure of human D-amino acid oxidase: context-dependent variability of the backbone conformation of the VAAGL hydrophobic stretch located at the si-face of the flavin ring. Protein Sci 15, 2708-2717

31. Kawazoe, T., Park, H.K., Iwana, S., Tsuge, H., and Fukui, K. (2007) Human D-amino acid oxidase: an update and review. Chem Rec 7, 305-315

32. Kennaway, C.K., Obarska-Kosinska, A., White, J.H., Tuszynska, I., Cooper, L.P., Bujnicki, J.M., Trinick, J., and Dryden, D.T. (2009) The structure of M.EcoKI Type I DNA methyltransferase with a DNA mimic antirestriction protein. Nucleic Acids Res 37, 762-770

33. Tesmer, V.M., Kawano, T., Shankaranarayanan, A., Kozasa, T., and Tesmer, J.J. (2005) Snapshot of activated G proteins at the membrane: the Galphaq-GRK2-Gbetagamma complex. Science 310, 1686-1690

Table 1 Validation of the final model of ND

Ramachandran plot			Verify3D	ERRAT
Favored	Allowed	Outlier	%	
%	%	%		
95.7	4.3	0.0	94.29	100.000

Table 2 Validation of the final model of CD

Ramachandran plot			Verify3D	ERRAT
Favored	Allowed	Outlier	%	
%	%	%		
90.0	10.0	0.0	74.39	100.000

Figure legends

Fig. 1 the hybrid *ab initio* approach

(A) Workflow of the hybrid *ab initio* approach. Domains were defined according to Conserved Domain Database. After producing models/decoys, two rounds of clustering were performed to find the consensus cluster. The model with the most average structure was selected as the best model before the MD analysis for the refinement and stability check. Finally the structure was evaluated with the Ramachandran plot, Verify3D and ERRAT. (B) Evaluation of the hybrid *ab initio* approach. The sequences of six PDB coordinates were used to test the hybrid approach. The accuracy of the approach was evaluated with the RMSD between the full-length predicted structure and experimentally solved structure (RMSD-a). RMSD-b indicates the values that were calculated by excluding 10 residues from N- and C-termini in each comparison. (C) A visual comparison of experimentally solved (*left*, PDB code: 3FEA) and predicted structures (*right*).

Fig. 2 Domain definition of G72

According to Conserved Domain Database (*23*), a domain is defined from residue 72 to 153 of the 153 residue-long isoform of G72. We termed this domain CD. The remaining region (residue 1 – 71) was termed ND.

Fig. 3 Process of the selection of the best model of ND

(A) Clustering process of the ND models/decoys. 22000 decoys from AbinitioRelax (AR), 5 models from I-TASSER (IT) and 10 models from Quark (Q) were clustered. The clusters colored blue were the consensus clusters that contained the models/decoys from all the three algorithms.

(B) Superposition of the models/decoys from AbinitioRelax (beige), I-TASSER (sky blue), Quark (magenta) in one of the consensus clusters. The location of Arg30 is shown in green.

(C) The final comparisons of the models/decoys shown in (B) based on RMSD. The compared models were the 1st model from I-TASSER (IT1), 4th model from QUARK (Q4) and model c.0.3 from AbinitioRelax (AR c.0.3).

Fig. 4 Final comparisons of models/decoys of CD

(A) Superposition of the models/decoys from AbinitioRelax (beige), I-TASSER (sky blue), Quark (magenta) in one of the consensus clusters.

(B) The final comparison of the models/decoys shown in (A) based on RMSD. The compared models were the 2nd model from I-TASSER (IT2), 8th model from QUARK (Q8) and model c.0.4 from AbinitioRelax (AR c.0.4).

Fig. 5 Surface charge distribution of ND and CD

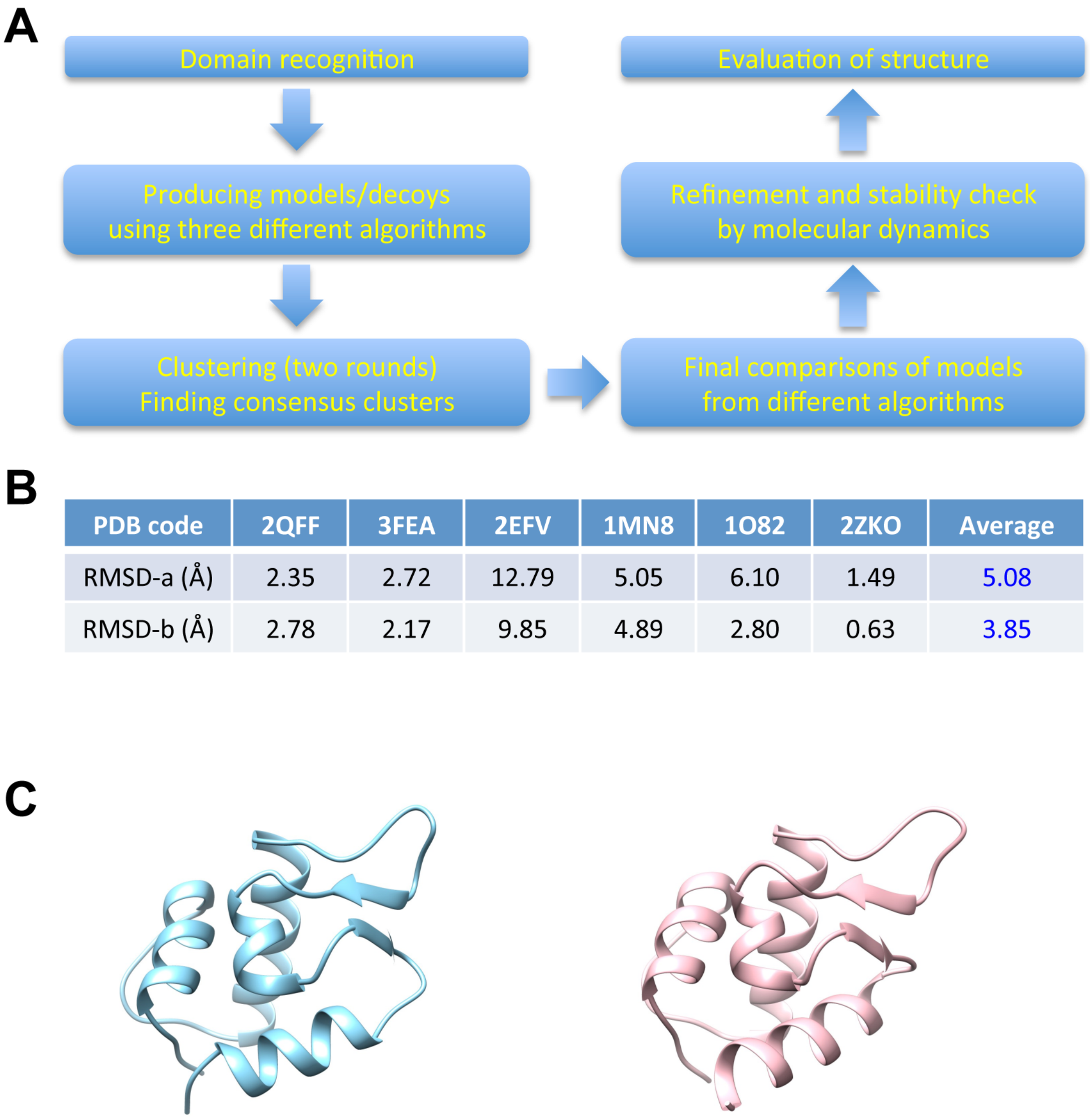
Positive and negative surface charges are colored blue and red, respectively. (A) Front view of the best ND model with surface representation. Positive residues that surround Arg30 are indicated. (B) Back view of ND. (C) Front view of the best CD model. (D) Back view of CD.

Fig. 6 Fold search of ND

(A - C) Superposition of the ND model on 4X82 (A), 5AEZ (B) and 1QQ0 (C) that were searched by COFACTOR. 4X82 (Rank1), 5AEZ (Rank3) and 1QQ0 (Rank6) are the PDB codes of the extracellular domain of ZIP4, Mep2 ammonium transceptor and carbonic anhydrase, respectively. ND was colored sky blue, whereas the other proteins are colored gray. Shade on the protein structure in (B) indicates the thickness of a lipid bilayer. (D) The structurally aligned region of OGT (PDB code; 2XGO) with ND by the Dali server. (E) The ND model of G72.

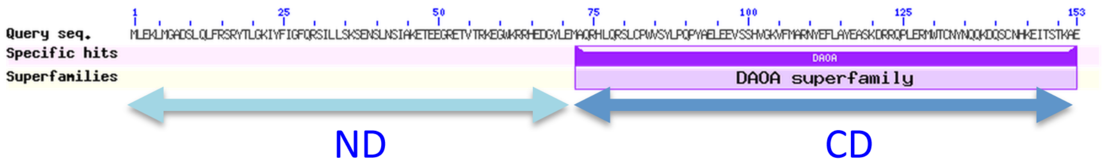
Fig. 7 Fold search of CD

(A - C) Superposition of CD (sky blue) on 2OKC (A), 2OAB (B) and 1G7V (C) searched by COFACTOR. 2OKC (gray), 2OAB (gray) and 1G7V (gray) are the PDB codes of HsdM, 3-deoxy-D-arabino-heptulosonate-7-phosphate synthase and 2-dehydro-3-deoxyphosphooctonate aldolase, respectively. (D) The CD model (E) The structure of a part of HsdM (yellow) is aligned with that of CD in (D). The yellow and gray ribbon models form the interface of the homodimer (PDB code: 2Y7C). (F) The RH bundle subdomain of GRK2 (yellow) interacting with Gαq (pink). PDB code is 2BCJ. The RH bundle subdomain is depicted in the same orientation as CD.


A
Domain recognition
Producing models/decoys using three different algorithms
Clustering (two rounds) Finding consensus clusters
Final comparisons of models from different algorithms
Refinement and stability check by molecular dynamics
Evaluation of structure
B
C

PDB code	2QFF	3FEA	2EFV	1MN8	1O82	2ZKO	Average
RMSD-a (Å)	2.35	2.72	12.79	5.05	6.10	1.49	5.08
RMSD-b (Å)	2.78	2.17	9.85	4.89	2.80	0.63	3.85

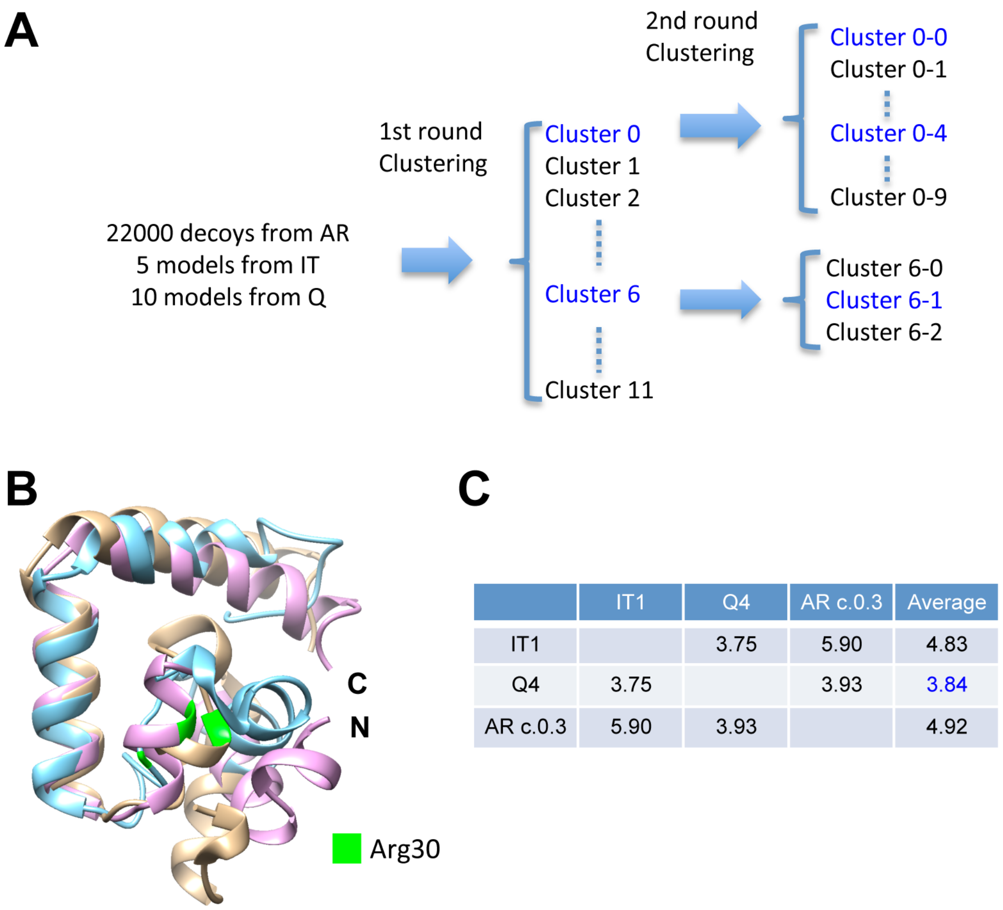
Fig. 1 Kato and Fukui



```
Query seq.      MLEKLMGADSLQLFRSRYTLGKIYFIGFQRSILLSKSENSLNSIAKETEEGRETVTRKEGWKRRHEDGYLEMAQRHLQRSLCPWVSYLPQPYAELEEVSSHVGKVFMARNYEFLAYEASKDRRQPLERMWTCNYNQQKDQSCNHKEITSTKAE
Specific hits   DAOA
Superfamilies   DAOA superfamily
```

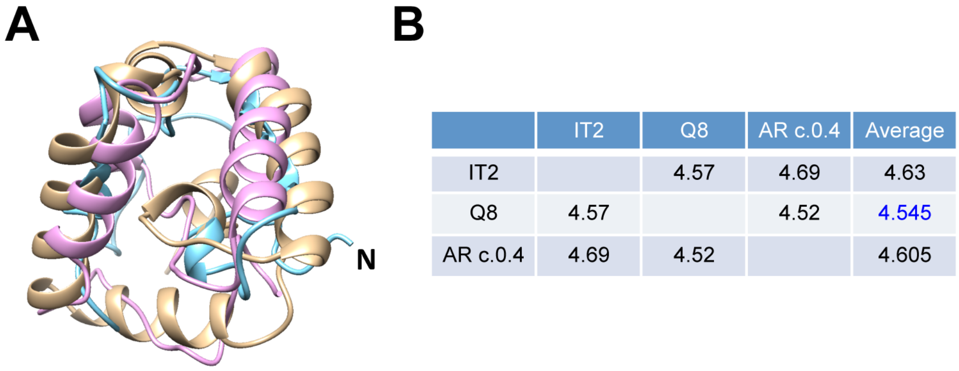
Fig. 2 Kato and Fukui

Domain definition



	IT1	Q4	AR c.0.3	Average
IT1		3.75	5.90	4.83
Q4	3.75		3.93	3.84
AR c.0.3	5.90	3.93		4.92

Fig. 3 Kato and Fukui



	IT2	Q8	AR c.0.4	Average
IT2		4.57	4.69	4.63
Q8	4.57		4.52	4.545
AR c.0.4	4.69	4.52		4.605

Fig. 4 Kato and Fukui

Final Comparison of model structures (CD)

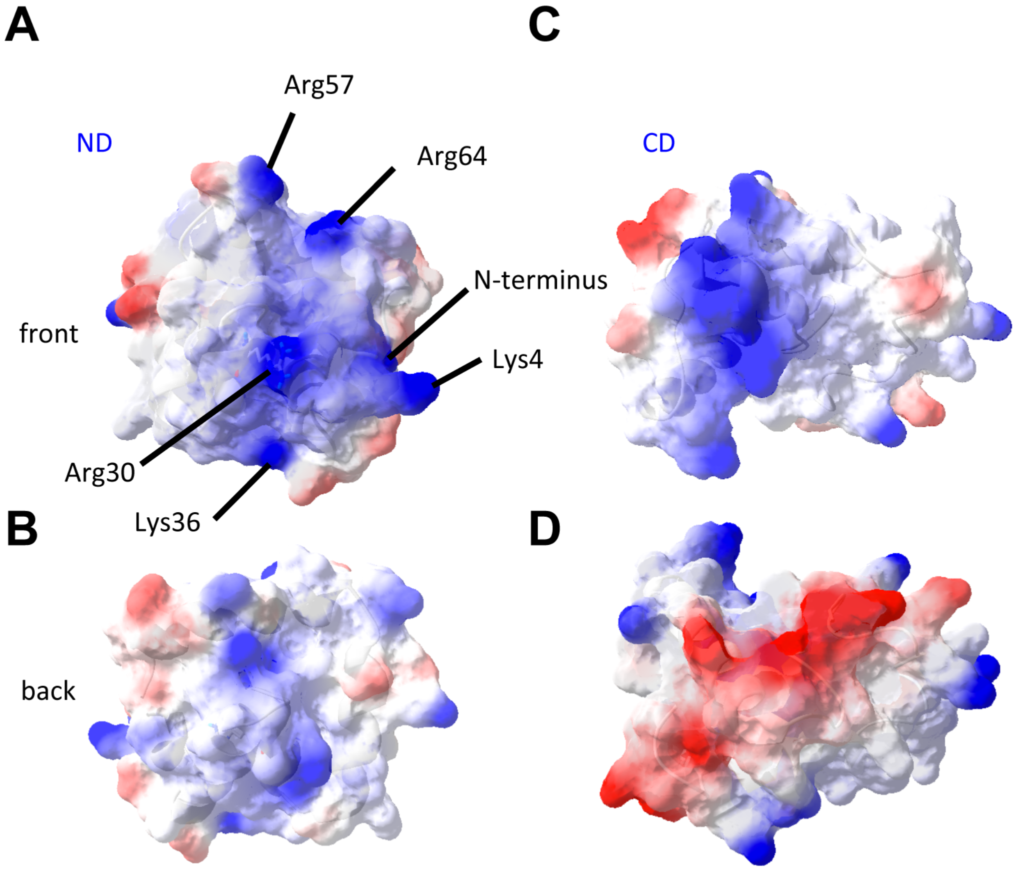


Fig. 5 Kato and Fukui

Surface charge distribution (ND&CD)

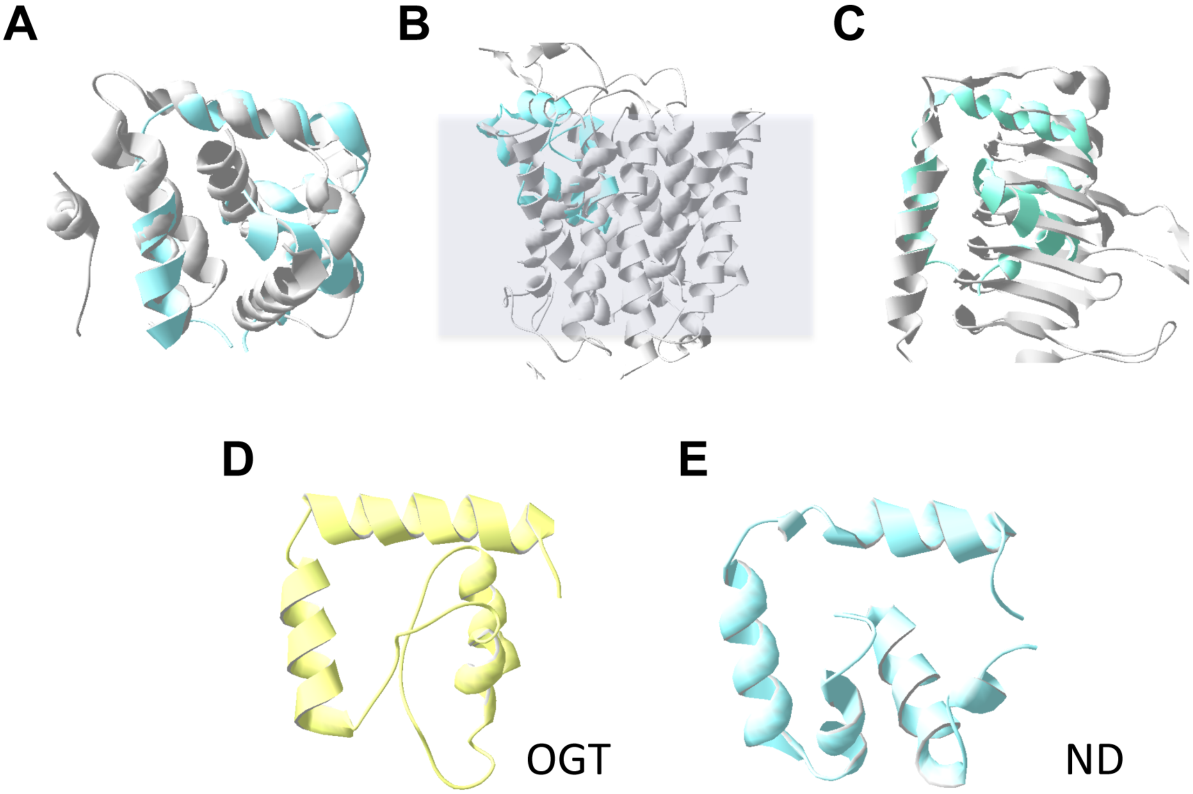


Fig. 6 Kato and Fukui

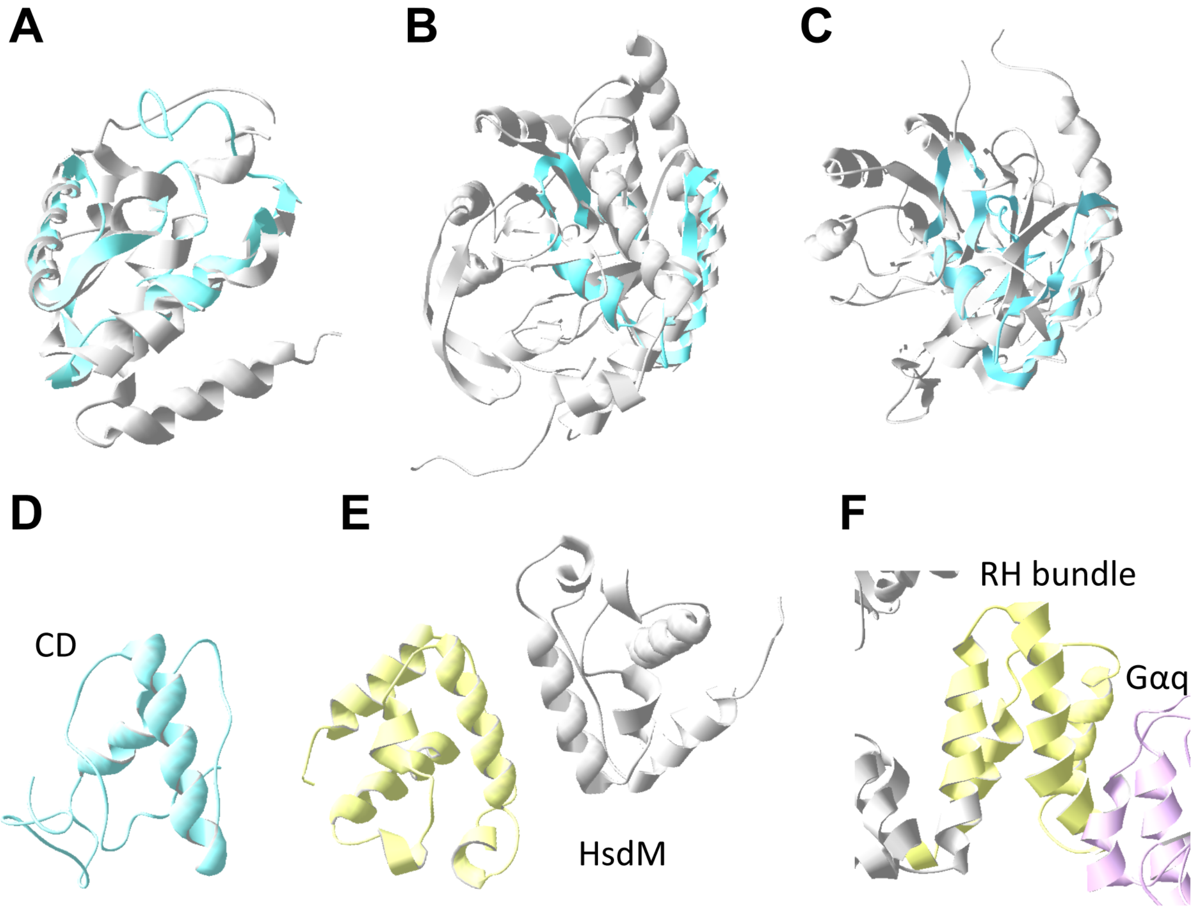


Fig. 7 Kato and Fukui

Supplementary Table S1

Top 10 identified structural analogs of ND in PDB analyzed by COFACTOR

Rank	PDB Hit	TM-score[a]	RMSD[b]	Identity[c]	Coverage[d]
1	4x82B	0.564	3.32	0.057	0.944
2	3nyyA	0.551	3.18	0.015	0.915
3	5aezA	0.537	3.36	0.043	0.873
4	5af3A	0.535	3.09	0.056	0.845
5	3c8vC	0.530	3.38	0.092	0.873
6	1qq0A	0.521	3.77	0.030	0.901
7	3c8iB	0.520	3.41	0.081	0.859
8	3r1wA	0.517	3.57	0.030	0.901
9	3i5oB	0.517	2.98	0.000	0.789
10	2y35A	0.516	3.29	0.059	0.873

[a] TM-score of the structural alignment between the query structure and known structures in the PDB library.

[b] RMSD between residues that are structurally aligned by TM-align.

[c] The sequence identity in the structurally aligned region.

[d] Coverage represents the coverage of the alignment by TM-align and is equal to the number of structurally aligned residues divided by length of the query protein.

Supplementary Table S2

Structural analogs of ND in PDB analyzed by the Dali server

No	Chain	Z[a]	rmsd	lali[b]	nres[c]	%id[d]	Description
1	2xgo-B	3.2	7	54	541	6	XCOGT;
2	2xgs-A	3.2	7	54	542	6	XCOGT;
3	2xgm-B	3.1	7.1	54	542	6	XCOGT;
4	2vsy-B	3.1	7.1	54	547	6	XCC0866;
5	2jlb-B	3.1	7.1	54	548	6	XCC0866;

6	1pjt-A	3	3	49	449	10	SIROHEME SYNTHASE;
7	1pjq-A	2.9	3	49	448	10	SIROHEME SYNTHASE;
8	1pjs-A	2.9	3	49	444	10	SIROHEME SYNTHASE;
9	5a01-A	2.7	7.4	53	681	8	O-GLYCOSYLTRANSFERASE;
10	4gyw-C	2.6	6.9	54	674	13	UDP-N-ACETYLGLUCOSAMINE--PEPTIDE N-
11	4xif-A	2.6	7.2	53	702	13	UDP-N-ACETYLGLUCOSAMINE--PEPTIDE N-
12	4gz3-C	2.6	6.9	54	674	13	UDP-N-ACETYLGLUCOSAMINE--PEPTIDE N-
13	3pe4-C	2.6	6.9	54	674	13	UDP-N-ACETYLGLUCOSAMINE--PEPTIDE N-
14	3pe3-A	2.6	7.2	53	701	13	UDP-N-ACETYLGLUCOSAMINE--PEPTIDE N-
15	4xi9-A	2.6	6.9	54	698	13	UDP-N-ACETYLGLUCOSAMINE--PEPTIDE N-
16	5a01-B	2.6	7.4	53	681	8	O-GLYCOSYLTRANSFERASE;
17	4cdr-D	2.6	6.9	54	698	13	UDP-N-ACETYLGLUCOSAMINE--PEPTIDE
18	4xi9-C	2.6	6.9	54	698	13	UDP-N-ACETYLGLUCOSAMINE--PEPTIDE N-
19	4cdr-B	2.6	6.9	54	698	13	UDP-N-ACETYLGLUCOSAMINE--PEPTIDE
20	3tax-A	2.6	6.9	54	695	13	UDP-N-ACETYLGLUCOSAMINE--PEPTIDE N-
21	4n3b-A	2.6	6.9	54	697	13	UDP-N-ACETYLGLUCOSAMINE--PEPTIDE N-
22	5bnw-A	2.6	6.9	54	694	13	UDP-N-ACETYLGLUCOSAMINE--PEPTIDE N-
23	3pe3-C	2.6	6.9	54	701	13	UDP-N-ACETYLGLUCOSAMINE--PEPTIDE N-
24	4xif-C	2.6	7.2	53	702	13	UDP-N-ACETYLGLUCOSAMINE--PEPTIDE N-
25	3tax-C	2.6	6.9	54	695	13	UDP-N-ACETYLGLUCOSAMINE--PEPTIDE N-
26	5c1d-A	2.6	6.9	54	695	13	UDP-N-ACETYLGLUCOSAMINE--PEPTIDE N-
27	4gz5-C	2.6	6.9	54	700	13	UDP-N-ACETYLGLUCOSAMINE--PEPTIDE N-
28	4ay6-A	2.6	7.2	53	698	13	UDP-N-ACETYLGLUCOSAMINE--PEPTIDE
29	4ay6-D	2.6	6.9	54	698	13	UDP-N-ACETYLGLUCOSAMINE--PEPTIDE
30	4ay5-A	2.6	6.9	54	698	13	UDP-N-ACETYLGLUCOSAMINE--PEPTIDE N-ACETYLGLUCOSAM
31	4ay5-B	2.6	6.9	54	698	13	UDP-N-ACETYLGLUCOSAMINE--PEPTIDE N-ACETYLGLUCOSAM
32	4gz3-A	2.6	6.9	54	695	13	UDP-N-ACETYLGLUCOSAMINE--PEPTIDE N-
33	4cdr-C	2.6	6.9	54	698	13	UDP-N-ACETYLGLUCOSAMINE--PEPTIDE
34	4xif-B	2.6	7.2	53	702	13	UDP-N-ACETYLGLUCOSAMINE--PEPTIDE N-
35	4gyw-A	2.6	6.9	54	695	13	UDP-N-ACETYLGLUCOSAMINE--PEPTIDE N-
36	4xi9-D	2.6	6.9	54	698	13	UDP-N-ACETYLGLUCOSAMINE--PEPTIDE N-

37	4cdr-A	2.6	6.9	54	698	13	UDP-N-ACETYLGLUCOSAMINE--PEPTIDE
38	4xif-D	2.6	7.4	53	702	13	UDP-N-ACETYLGLUCOSAMINE--PEPTIDE N-
39	3pe4-A	2.6	6.9	54	695	13	UDP-N-ACETYLGLUCOSAMINE--PEPTIDE N-
40	3pe3-D	2.6	6.9	54	701	13	UDP-N-ACETYLGLUCOSAMINE--PEPTIDE N-
41	3pe3-B	2.6	6.9	54	701	13	UDP-N-ACETYLGLUCOSAMINE--PEPTIDE N-
42	4n3a-A	2.6	7.2	53	697	13	UDP-N-ACETYLGLUCOSAMINE--PEPTIDE N-
43	4ay5-C	2.6	6.9	54	698	13	UDP-N-ACETYLGLUCOSAMINE--PEPTIDE N-ACETYLGLUCOSAM
44	4ay6-C	2.6	6.9	54	698	13	UDP-N-ACETYLGLUCOSAMINE--PEPTIDE
45	4gz6-A	2.6	6.9	54	700	13	UDP-N-ACETYLGLUCOSAMINE--PEPTIDE N-
46	4gz6-C	2.6	7	52	700	13	UDP-N-ACETYLGLUCOSAMINE--PEPTIDE N-
47	5a01-C	2.6	7.4	53	681	8	O-GLYCOSYLTRANSFERASE;
48	4xi9-B	2.6	6.9	54	698	13	UDP-N-ACETYLGLUCOSAMINE--PEPTIDE N-
49	4n3c-A	2.6	6.9	54	697	13	UDP-N-ACETYLGLUCOSAMINE--PEPTIDE N-
50	4gyy-A	2.6	6.9	54	693	13	UDP-N-ACETYLGLUCOSAMINE--PEPTIDE N-
51	4n39-A	2.6	6.9	54	697	13	UDP-N-ACETYLGLUCOSAMINE--PEPTIDE N-
52	4gz6-B	2.6	6.9	54	700	13	UDP-N-ACETYLGLUCOSAMINE--PEPTIDE N-
53	4gyy-C	2.6	6.9	54	671	13	UDP-N-ACETYLGLUCOSAMINE--PEPTIDE N-
54	4ay5-D	2.6	6.9	54	698	13	UDP-N-ACETYLGLUCOSAMINE--PEPTIDE N-ACETYLGLUCOSAM
55	4ay6-B	2.6	6.9	54	698	13	UDP-N-ACETYLGLUCOSAMINE--PEPTIDE
56	2jlb-A	2.5	7	54	548	6	XCC0866;
57	5djs-B	2.5	7	53	520	15	TETRATRICOPEPTIDE TPR_2 REPEAT PROTEIN;
58	5djs-A	2.5	7	53	520	15	TETRATRICOPEPTIDE TPR_2 REPEAT PROTEIN;
59	5djs-C	2.5	7	53	520	15	TETRATRICOPEPTIDE TPR_2 REPEAT PROTEIN;
60	2xgo-A	2.5	7.1	52	548	6	XCOGT;
61	2vsy-A	2.5	7.1	54	547	6	XCC0866;
62	2vsn-B	2.5	7.1	52	534	6	XCOGT;
63	4gz6-D	2.5	7	52	700	13	UDP-N-ACETYLGLUCOSAMINE--PEPTIDE N-
64	4gz5-A	2.5	6.9	54	700	13	UDP-N-ACETYLGLUCOSAMINE--PEPTIDE N-
65	4gz5-B	2.5	6.9	54	700	13	UDP-N-ACETYLGLUCOSAMINE--PEPTIDE N-
66	4gz5-D	2.5	6.9	54	700	13	UDP-N-ACETYLGLUCOSAMINE--PEPTIDE N-
67	2xgm-A	2.5	7.2	52	512	6	XCOGT;
68	2xa2-A	2.4	6.1	50	412	6	TREHALOSE-SYNTHASE TRET;

69	2vsn-A	2.4	7.1	52	534	6	XCOGT;
70	2bnk-B	2.4	3.1	46	64	7	EARLY PROTEIN GP16.7;
71	2c5r-F	2.4	3.3	47	63	6	EARLY PROTEIN P16.7;
72	2c5r-B	2.4	3.3	47	63	6	EARLY PROTEIN P16.7;
73	1zae-A	2.3	2.7	47	70	4	EARLY PROTEIN GP16.7;
74	4n9w-A	2.3	6.2	52	360	12	GDP-MANNOSE-DEPENDENT ALPHA-(1-2)-PHOSPHATIDYLINO
75	2c5r-E	2.3	3.3	47	63	6	EARLY PROTEIN P16.7;
76	2c5r-D	2.3	3.3	47	63	6	EARLY PROTEIN P16.7;
77	2bnk-A	2.3	3	45	64	7	EARLY PROTEIN GP16.7;
78	2c5r-C	2.3	3.3	47	63	6	EARLY PROTEIN P16.7;
79	2c5r-A	2.3	3.3	47	63	6	EARLY PROTEIN P16.7;
80	1pjt-B	2.3	3.5	51	456	10	SIROHEME SYNTHASE;
81	2x6q-B	2.3	6.2	49	409	6	TREHALOSE-SYNTHASE TRET;
82	5hes-A	2.2	2.9	55	288	5	MITOGEN-ACTIVATED PROTEIN KINASE KINASE KINASE ML
83	4ae4-B	2.2	3.1	51	115	10	UBIQUITIN-ASSOCIATED PROTEIN 1;
84	2dah-A	2.2	2.9	43	54	7	UBIQUILIN-3;
85	1pjq-B	2.2	3.2	48	456	13	SIROHEME SYNTHASE;
86	3q3e-B	2.2	7.4	50	596	4	HMW1C-LIKE GLYCOSYLTRANSFERASE;
87	2xmp-A	2.2	6.1	50	412	6	TREHALOSE-SYNTHASE TRET;
88	1zae-B	2.2	3.5	48	70	6	EARLY PROTEIN GP16.7;
89	3g6i-A	2.2	5.3	57	200	4	PUTATIVE OUTER MEMBRANE PROTEIN, PART OF CARBOHYD
90	1pjs-B	2.2	3.5	51	455	10	SIROHEME SYNTHASE;
91	3fx3-B	2.1	4.1	57	231	9	CYCLIC NUCLEOTIDE-BINDING PROTEIN;
92	2x49-A	2.1	3.1	50	333	4	INVASION PROTEIN INVA;
93	2fgy-A	2.1	3.1	45	471	4	CARBOXYSOME SHELL POLYPEPTIDE;
94	3q3i-A	2.1	7.4	52	620	4	HMW1C-LIKE GLYCOSYLTRANSFERASE;
95	3q3h-A	2.1	7.5	50	620	4	HMW1C-LIKE GLYCOSYLTRANSFERASE;
96	3q3e-A	2.1	7	52	620	4	HMW1C-LIKE GLYCOSYLTRANSFERASE;
97	4x7m-A	2.1	6	52	493	6	UNCHARACTERIZED PROTEIN;
98	4x6l-C	2.1	6	52	493	6	TARM;
99	4x6l-B	2.1	6	52	493	6	TARM;

100	5jem-F	2.1	5.3	37	42	5	INTERFERON REGULATORY FACTOR 3;
101	2bwe-G	2.1	3.1	44	46	9	DSK2;
102	2bwe-Q	2.1	3.1	44	47	9	DSK2;
103	2bwe-R	2.1	3.2	44	46	9	DSK2;
104	1l8y-A	2	3	45	83	7	UPSTREAM BINDING FACTOR 1;
105	3q3h-B	2	7	52	595	4	HMW1C-LIKE GLYCOSYLTRANSFERASE;
106	3q3i-B	2	7.4	50	593	4	HMW1C-LIKE GLYCOSYLTRANSFERASE;
107	3s28-E	2	5.8	52	781	4	SUCROSE SYNTHASE 1;
108	1wgn-A	2	3.6	46	63	7	UBIQUITIN ASSOCIATED PROTEIN;
109	5cra-A	2	5.2	54	172	11	SDEA;
110	4un2-B	2	2.7	39	43	13	UBIQUITIN;

[a] Z-score of the structural alignment between the query structure and known structures in the PDB library.

[b] Length of the alignment between the query structure and known structures.

[c] Number of aligned residues.

[d] The sequence identity (%) in the structurally aligned region.

Supplementary Table S3

Top 10 identified structural analogs of CD in PDB analyzed by COFACTOR

Rank	PDB Hit	TM-score [a]	RMSD [b]	Identity [c]	Coverage [d]
1	2okcB1	0.560	2.53	0.048	0.768
2	1oabB	0.544	3.83	0.049	0.927
3	1g7vA	0.539	3.84	0.038	0.890
4	3fs2B	0.538	3.87	0.049	0.915
5	2nrjA	0.534	3.38	0.038	0.829
6	4k82A	0.533	3.91	0.050	0.939
7	3tmqA	0.530	4.00	0.062	0.915
8	4ur5A	0.526	3.88	0.104	0.915
9	3t4cA	0.526	3.87	0.061	0.866

10	3e9aA	0.516	3.66	0.049	0.878

[a] TM-score of the structural alignment between the query structure and known structures in the PDB library.

[b] RMSD between residues that are structurally aligned by TM-align.

[c] The sequence identity in the structurally aligned region.

[d] Coverage represents the coverage of the alignment by TM-align and is equal to the number of structurally aligned residues divided by length of the query protein.

Supplementary Table S4

Structural analogs of CD in PDB analyzed by the Dali server

No	Chain	Z[a]	rmsd	lali[b]	nres[c]	%id[d]	Description
1	3uzt-A	2.9	3.3	56	586	7	BETA-ADRENERGIC RECEPTOR KINASE 1;
2	3pvu-A	2.7	3.3	57	609	9	BETA-ADRENERGIC RECEPTOR KINASE 1;
3	2bcj-A	2.6	3.4	57	624	9	G-PROTEIN COUPLED RECEPTOR KINASE 2;
4	1ym7-D	2.6	3.4	57	599	9	BETA-ADRENERGIC RECEPTOR KINASE 1;
5	2acx-B	2.6	3.5	58	492	5	G PROTEIN-COUPLED RECEPTOR KINASE 6;
6	3nyo-A	2.6	3.3	58	553	5	G PROTEIN-COUPLED RECEPTOR KINASE 6;
7	2bv1-A	2.5	3.2	56	134	14	REGULATOR OF G-PROTEIN SIGNALLING 1;
8	5do9-F	2.5	3.8	56	134	13	GUANINE NUCLEOTIDE-BINDING PROTEIN G(Q) SUBUNIT A
9	1ym7-A	2.5	3.4	57	608	9	BETA-ADRENERGIC RECEPTOR KINASE 1;
10	1ym7-B	2.5	3.4	57	608	9	BETA-ADRENERGIC RECEPTOR KINASE 1;
11	3nyn-B	2.5	3.4	57	553	5	G PROTEIN-COUPLED RECEPTOR KINASE 6;
12	5do9-D	2.4	3.4	54	134	15	GUANINE NUCLEOTIDE-BINDING PROTEIN G(Q) SUBUNIT A
13	5do9-B	2.4	3.8	56	134	13	GUANINE NUCLEOTIDE-BINDING PROTEIN G(Q) SUBUNIT A
14	3v5w-A	2.4	3.3	56	623	5	G-PROTEIN COUPLED RECEPTOR KINASE 2;
15	3nyn-A	2.4	3.4	58	553	5	G PROTEIN-COUPLED RECEPTOR KINASE 6;
16	2gtp-C	2.3	3.3	56	132	14	GUANINE NUCLEOTIDE-BINDING PROTEIN G(I), ALPHA-1
17	2gtp-D	2.3	3.3	56	132	14	GUANINE NUCLEOTIDE-BINDING PROTEIN G(I), ALPHA-1
18	3cik-A	2.3	3.4	58	619	7	BETA-ADRENERGIC RECEPTOR KINASE 1;

19	4amq-A	2.2	3.8	55	341	7	L544;
20	1emu-A	2.1	3.7	58	132	5	AXIN;
21	3c51-B	2.1	3.2	50	461	6	RHODOPSIN KINASE;
22	4ekd-B	2.1	3.1	56	132	9	GUANINE NUCLEOTIDE-BINDING PROTEIN G(Q) SUBUNIT A
23	1dk8-A	2	3.5	58	147	5	AXIN;
24	4ekc-D	2	3.2	56	128	9	GUANINE NUCLEOTIDE-BINDING PROTEIN G(Q) SUBUNIT A
25	4gou-A	2	3.4	56	507	13	EHRGS-RHOGEF;
26	3c4w-B	2	3.3	53	519	6	RHODOPSIN KINASE;

[a] Z-score of the structural alignment between the query structure and known structures in the PDB library.

[b] Length of the alignment between the query structure and known structures.

[c] Number of aligned residues.

[d] The sequence identity (%) in the structurally aligned region.

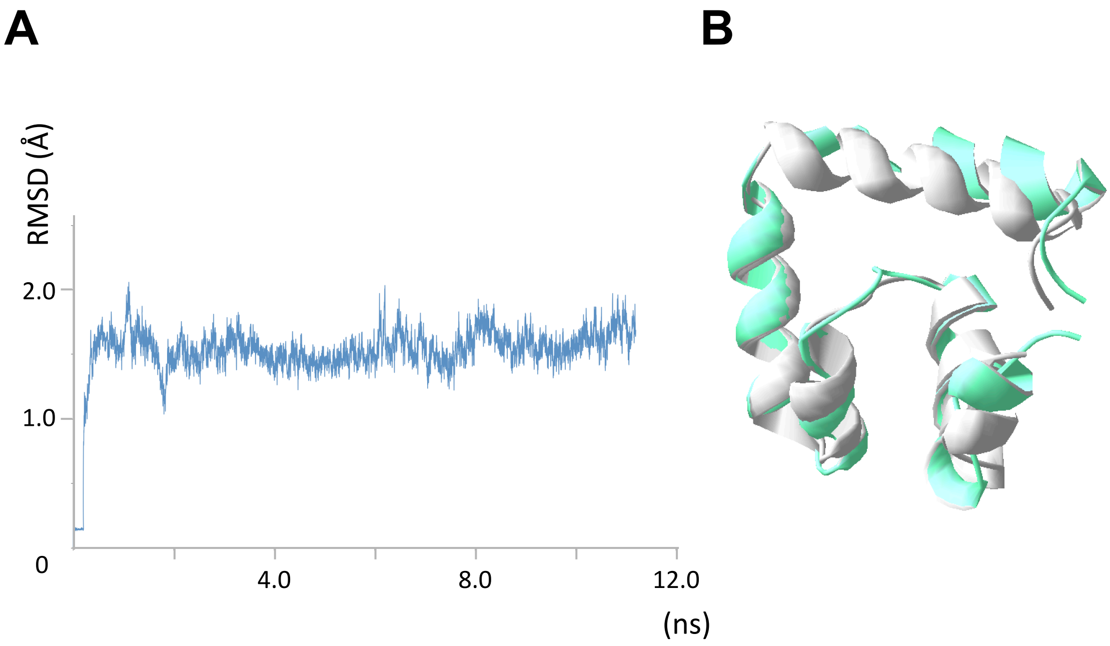
```
                                            30                            60
                MLEKLMGADSLQLFRSRYTLGKIYFIGFQRSILLSKSENSLNSIAKETEEGRETVTRKEG
JPRED           -HHHHH----HHHHHHHHHH--EEHH---HHHH-------HHHHH----------HH---
PSIPRED         -HHHHH--HHHHHHHHHHH---EEEEEEE--HHH----HHHHHHHHHHHH----HHHHHH

                                            90                           120
                WKRRHEDGYLEMAQRHLQRSLCPWVSYLPQPYAELEEVSSHVGKVFMARNYEFLAYEASK
JPRED           ---------HHHHHHHHHHH------------HHHHHHHHHHHHHHH----HEHHH----
PSIPRED         H------HHHHHHHHHHHHH------------HHHHHHHHHHHHHHHH--HHHHHHHHHH

                                           150
                DRRQPLERMWTCNYNQQKDQSCNHKEITSTKAE
JPRED           ----HHHHHHHH---------------------
PSIPRED         HH-----HHHH----------------------
```

This figure illustrates the secondary structure prediction within G72 by the JPRED and PSIPRED algorithms. H and E indicate α-helix and extended structure (β-strand), respectively.

Supplementary Fig. S1
Secondary structure prediction
Kato and Fukui

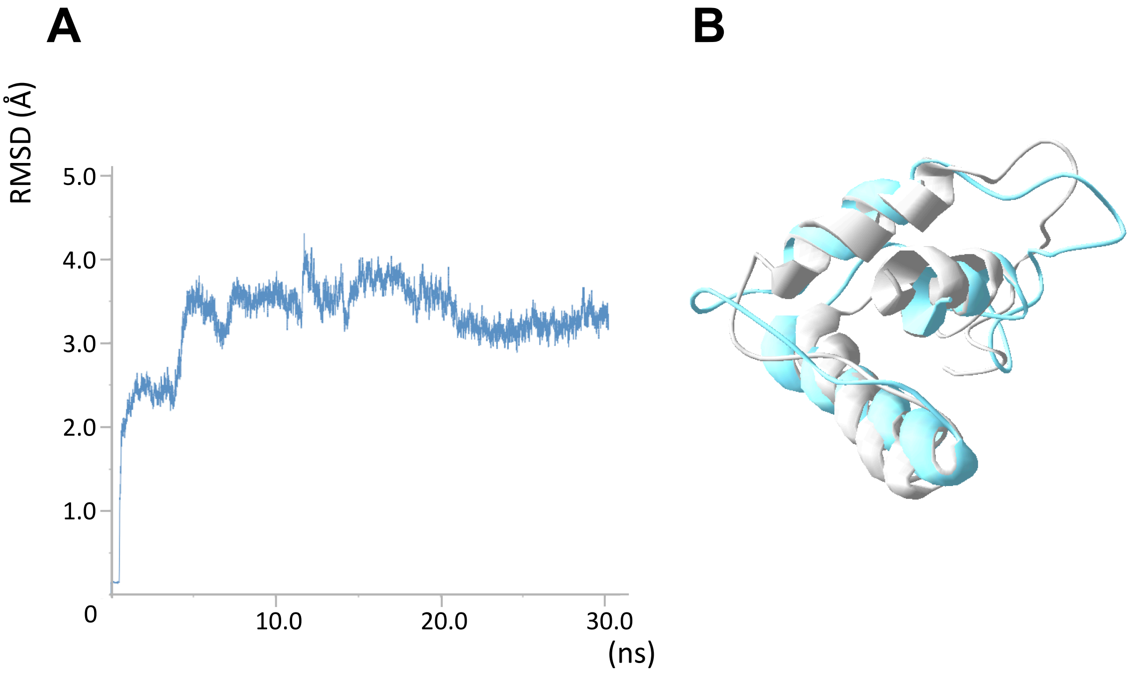


(A) Time course of the RMSD values between the initial structure and trajectories in the MD simulation.

(B) Superposition of the initial (gray) and final (green) trajectories in the MD simulation.

Supplementary Fig. S2
Results of the MD simulation of the ND model

Kato and Fukui



(A) Time course of the RMSD values between the initial structure and trajectories in the MD simulation.

(B) Superposition of the initial (gray) and final (sky blue) trajectories in the MD simulation.

Supplementary Fig. S3
Results of the MD simulation of the CD model

Kato and Fukui