scientific reports

OPEN

Solution structure of *Gaussia* Luciferase with five disulfide bonds and identification of a putative coelenterazine binding cavity by heteronuclear NMR

Nan Wu[1,6], Naohiro Kobayashi[2,6], Kengo Tsuda[3], Satoru Unzai[4], Tomonori Saotome[5], Yutaka Kuroda[5✉] & Toshio Yamazaki[2✉]

Gaussia luciferase (GLuc) is a small luciferase (18.2 kDa; 168 residues) and is thus attracting much attention as a reporter protein, but the lack of structural information is hampering further application. Here, we report the first solution structure of a fully active, recombinant GLuc determined by heteronuclear multidimensional NMR. We obtained a natively folded GLuc by bacterial expression and efficient refolding using a Solubility Enhancement Petide (SEP) tag. Almost perfect assignments of GLuc's ^1H, ^{13}C and ^{15}N backbone signals were obtained. GLuc structure was determined using CYANA, which automatically identified over 2500 NOEs of which > 570 were long-range. GLuc is an all-alpha-helix protein made of nine helices. The region spanning residues 10–18, 36–81, 96–145 and containing eight out of the nine helices was determined with a C_α-atom RMSD of 1.39 Å ± 0.39 Å. The structure of GLuc is novel and unique. Two homologous sequential repeats form two anti-parallel bundles made by 4 helices and tied together by three disulfide bonds. The N-terminal helix 1 is grabbed by these 4 helices. Further, we found a hydrophobic cavity where several residues responsible for bioluminescence were identified in previous mutational studies, and we thus hypothesize that this is a catalytic cavity, where the hydrophobic coelenterazine binds and the bioluminescence reaction takes place.

Luciferase (Luc) is a generic term for bioluminescent enzymes that catalyze the oxidation of a substrate, often termed luciferin ([1], pages xix-xxi). Together with GFP, Luc is widely employed as a reporter protein[2–4]. *Gaussia* Luciferase (GLuc) is a luciferase isolated from the marine *Gaussia princeps*[5], which catalyzes a bright blue light by oxidizing coelenterazine. GLuc is small and has a molecular mass of 18.2 kDa (excluding the secretion tag). Nonetheless, its bioluminescence intensity is strong (200 fold higher than Firefly Luciferase and *Renilla* Luciferase, the two most widely used luciferase), and it is thus considered as a potential ideal reporter protein[6]. Attempts to improve or redesign GLuc's bioluminescence characteristics included the lengthening of its half-life luminescence[7–9], and the redshift of its light emission peak at 480 nm[10,11], which is absorbed by tissues during in vivo applications[12]. However, structural information at atomic resolution is still not available, making the redesign process tedious.

GLuc contains 10 cysteines, and previous studies demonstrated that the natively folded GLuc contains five disulfide bonds. The presence of 5 disulfide bonds increases the risks of misfolding when GLuc is bacterially produced, resulting in a low yield[13]. In order to overcome this misfolding problem, several methods including

[1]College of Food and Bioengineering, Zhengzhou University of Light Industry, 136 Kexue Road, Zhengzhou 450001, People's Republic of China. [2]NMR Science and Development Division, RSC, RIKEN, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama City, Kanagawa 230-0045, Japan. [3]Division of Structural and Synthetic Biology, Center for Life Science Technologies, RIKEN, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama City, Kanagawa 230-0045, Japan. [4]Department of Frontier Bioscience, Faculty of Bioscience and Applied Chemistry, Hosei University, 3-7-2, Kajino-cho, Koganei-shi, Tokyo 184-8584, Japan. [5]Department of Biotechnology and Life Science, Graduate School of Engineering, Tokyo University of Agriculture and Technology (TUAT), 2-24-16 Nakamachi, Koganei-shi, Tokyo 184-8588, Japan. [6]These authors contributed equally: Nan Wu and Naohiro Kobayashi. ✉email: ykuroda@cc.tuat.ac.jp; toshio.yamazaki@riken.jp
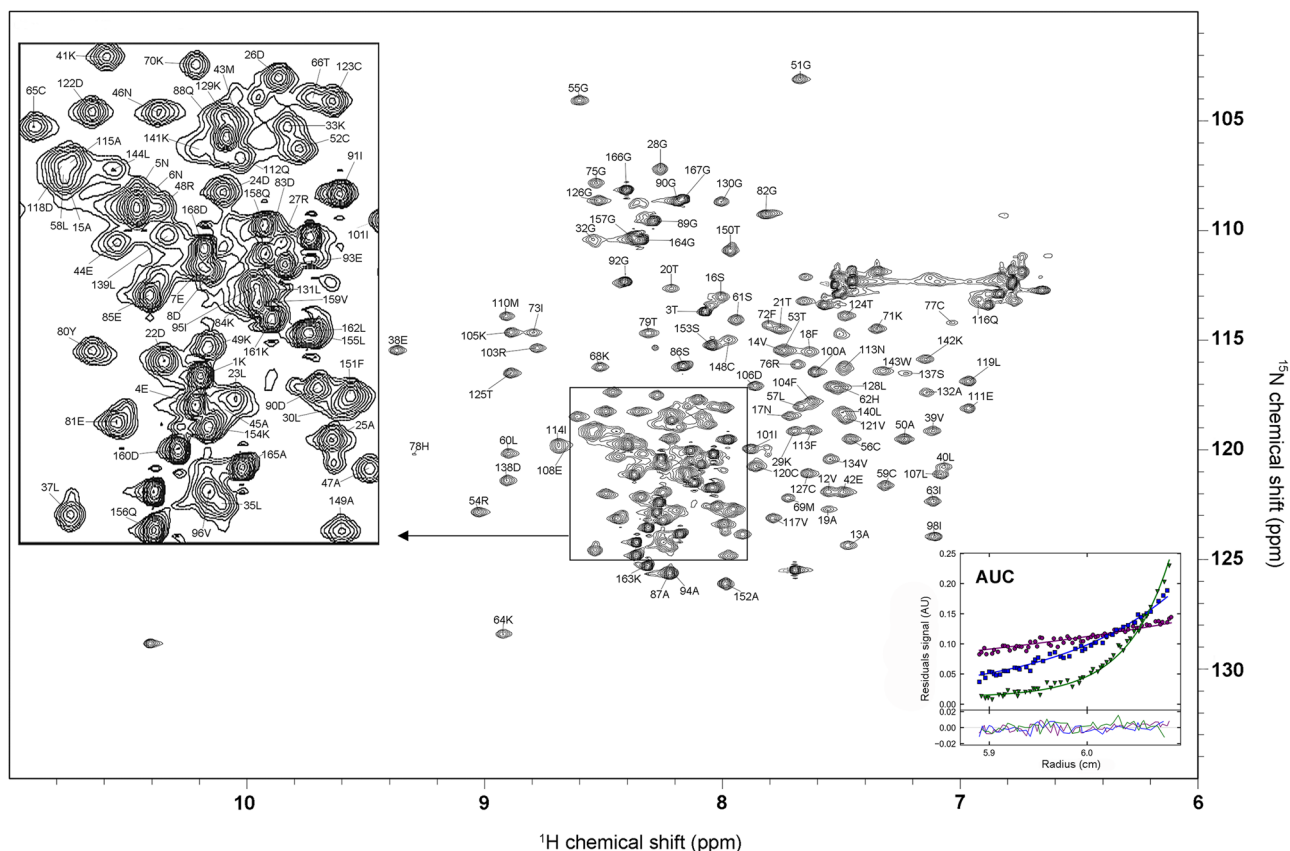
Figure 1. 2D ^{1}H–^{15}N HSQC spectrum of GLuc. The peak assignments are shown using the one-letter code followed by the residue number. Resonance assignments are numbered starting at the first residue (lysine) behind the secretion tag, which was removed without affecting the bioluminescence activity. Mutations at E100A and G103R do not affect activity and are described in our prior paper[13]. The inset in the right bottom (marked AUC) shows the result of sedimentation equilibrium experiments of GLuc protein (concentration at 0.3 mg/mL). Scans from three different rotor speeds (●: 12,000 rpm; ■: 22,000 rpm; ▼: 37,000 rpm) monitored at 280 nm. The lines represent the fit to a single species model. The determined molecular weight was 22 kDa, corresponding to a monomer.

fusion with pelB leader sequence[7,14], cell-free systems[15], low-temperature expression[16] were reported, but the yield of natively folded GLuc remained insufficient for high-resolution structural studies. We previously developed a Solubility Enhancement Peptide tag (SEP tag)[17–19]. We showed that by attaching a SEP tag containing nine aspartic acids to GLuc's C-terminus, we could increase the solubility of GLuc, resulting in a spontaneous refolding and the formation of native disulfide bonds. Indeed, we obtained nearly 1 mg of soluble and functional GLuc from a 200 ml of *E.coli* cultured in Luria–Bertani (LB)[11,13,20].

Here, we used the SEP-tag fused GLuc construct to produce a sufficient amount of ^{15}N and ^{13}C uniformly labeled GLuc for NMR studies. Heteronuclear multidimensional NMR spectroscopy enabled over 99% backbone ^{1}H, ^{13}C, and ^{15}N chemical shifts of GLuc to be assigned. Flexible regions and highly stable regions were identified by ^{1}H–^{15}N heteronuclear NOE[21] and H/D exchange experiments[22]. The three-dimensional structure calculated by using CYANA (ver 3.98[23]) were determined with a backbone ($C_α$) RMSD of 1.39Å ± 0.39Å (excluding residues in the flexible regions).

Results

Expression and purification of GLuc. The natively folded GLuc possesses ten cysteines that form five disulfide bonds, which can be easily misformed when the protein is expressed in *E.coli*, and the cysteines are air-oxidized in vitro. Here, we used a SEP-Tag, C9D, which solubilizes the protein during air-oxidization and refolding, thereby increasing the yield of natively folded and active GLuc[20]. The final yield of GLuc after tags cleavage and two times HPLC purification (Supplementary Fig. 1) was 1.5 mg per liter of M9 minimal medium culture, which was sufficient for NMR analysis. GLuc's identity and folding were confirmed by, respectively, MALDI-TOF mass (^{15}N labeled GLuc, calculated = 19,055.8 Da, experimental = 19,062.5 Da, see Supplementary Fig. 2) and bioluminescence activity measurement. The ^{15}N labeled GLuc's activity was essentially identical to that given in our previous report[11,13] (Supplementary Fig. 3). To date, the yield of natively folded active GLuc is almost nil when expressed without the C9D tag[20], and the solubilization tag was thus essential to achieve the present amount of protein, though it was removed once the protein was folded into its native conformation.

Residue	R2(50 Hz)-R2(1 kHz) 1/s
V12	6.66
S16	4.84
T21	2.47
D26	2.12
G28	5.60
L37	2.88
A47	2.61
S61	6.93
M69	4.87
K70	2.07
G75	4.78
T79	6.43
T125	2.26
D138	4.46
L140	4.07
T150	**4.46**
A152	**10.85**

Table 1. R_2-dispersion experiments on GLuc. Differences of effective ^{15}N transverse relaxation rates at CPMG rates of 50 Hz and 1 kHz are listed only for peaks that are resolved and their rate difference > 2 1/s. Residues that contribute to the central cavity were underlined. Larger difference was observed for residues in the C-terminal flexible region (indicated by bold letters) indicative of its structural dynamics. The two-state model analysis of CPMG rates of 50, 100, 150, 200, 250, 300, 400, 500, 600, 800, 1000 Hz showed that the estimated exchange rate was 2500 ± 800 1/s. Because the residues showing R2 dispersion spread around several blocks, we judged farther residue-specific analysis using single exchange rate is unreliable.

NMR analysis. The ^{1}H–^{15}N HSQC spectrum exhibited dispersed and sharp peaks (Fig. 1), indicating a stable and well-folded structure. Almost all backbone chemical shifts were visible in the heteronuclear NMR experiments, and over 99% of backbone ^{1}H, ^{13}C and ^{15}N resonances of non-proline residues were unambiguously assigned. C136 was the only un-assigned backbone H–N chemical shifts. The broadened signals around residue C136 suggested that the region encompassing the C136/C148 disulfide bond was subjected to structural exchange, as suggested by the ^{15}N relaxation dispersion of D138 and L140 (Table 1), and thus the C136 H–N pair was undetected.

The side-chain atoms were automatically assigned by FLYA[24] (a function of CYANA) using the aliphatic atoms identified in the 3D HCCH-TOCSY, ^{15}N- and ^{13}C-edited NOESY spectra. The assignments were confirmed by visual inspection and when necessary corrected manually using the NMR spectra viewer and analyzer MagRO[25,26]. We assigned over 82.4% of ^{1}H, ^{13}C and ^{15}N atoms of entire GLuc molecule.

The secondary structure elements were analyzed by TALOS + using the ^{1}H, ^{13}C, ^{15}N chemical shifts (Fig. 2a). TALOS + indicated that GLuc contains 36.9% helix and 4.7% sheets, in reasonable agreement with our previous prediction based on the consensus of seven publicly available secondary structure predictors (30% helix and 4% sheets) as well as with the results of our Circular Dichroism (CD) analysis (30% helix and 12% sheets)[13]. In addition, the location of helices calculated by TALOS + and the secondary structure prediction mostly overlapped (Supplementary Fig. 4).

Structure calculation and disulfide bond determination. Since GLuc was a monomer as demonstrated by AUC, all NOEs were used as intramolecular NOEs (Fig. 1). The statistics of NOEs assigned during the 19 CYANA runs are shown in Table 2. Even though distance constraints for hydrogen bonds, disulfide bonds, and some manually assigned NOEs were included in the CYANA calculations in addition to the standard automatically assigned NOEs, the target functions were reasonably small (4.07 ± 0.59), indicating that the resulting structures were consistent with the experimental data.

Ellman's assay indicated that all ten cysteines (C52, C56, C59, C65, C77, C120, C123, C127, C136, and C148) are oxidized in the active GLuc, and thus that they should form five disulfide bonds[20]. Three disulfide bonds C59/C120, C65/C77, and C136/C148 were unambiguously visible in the NMR structures, but the pairing of the remaining four cysteines (C52, C56, C123, and C127), which were close to each other, was less straightforward to determine. In order to determine the remaining two disulfide bonds, we set the distance between the gamma sulfur (S_γ) to > 2Å so that any cysteine could freely combine with any of the remaining three cysteines. We then identified cysteine pairs with S_γ distance < 3Å in the 380 structures obtained from 19 rounds. As a result, C52/C127 and C56/C123 were the most favored pairs and were observed in 92.4% and 56.1% of the calculated structures, respectively. On the other hand, C52/C56 and C123/C127 were observed in only 13.7% and 19.5% of the structures.

The disulfide bond combination between the cysteines (C52, C56, C59, C120, C123 and C127) were further characterized by limited proteolysis of GLuc with trypsin, and Liquid Chromatography–Mass Spectrometry
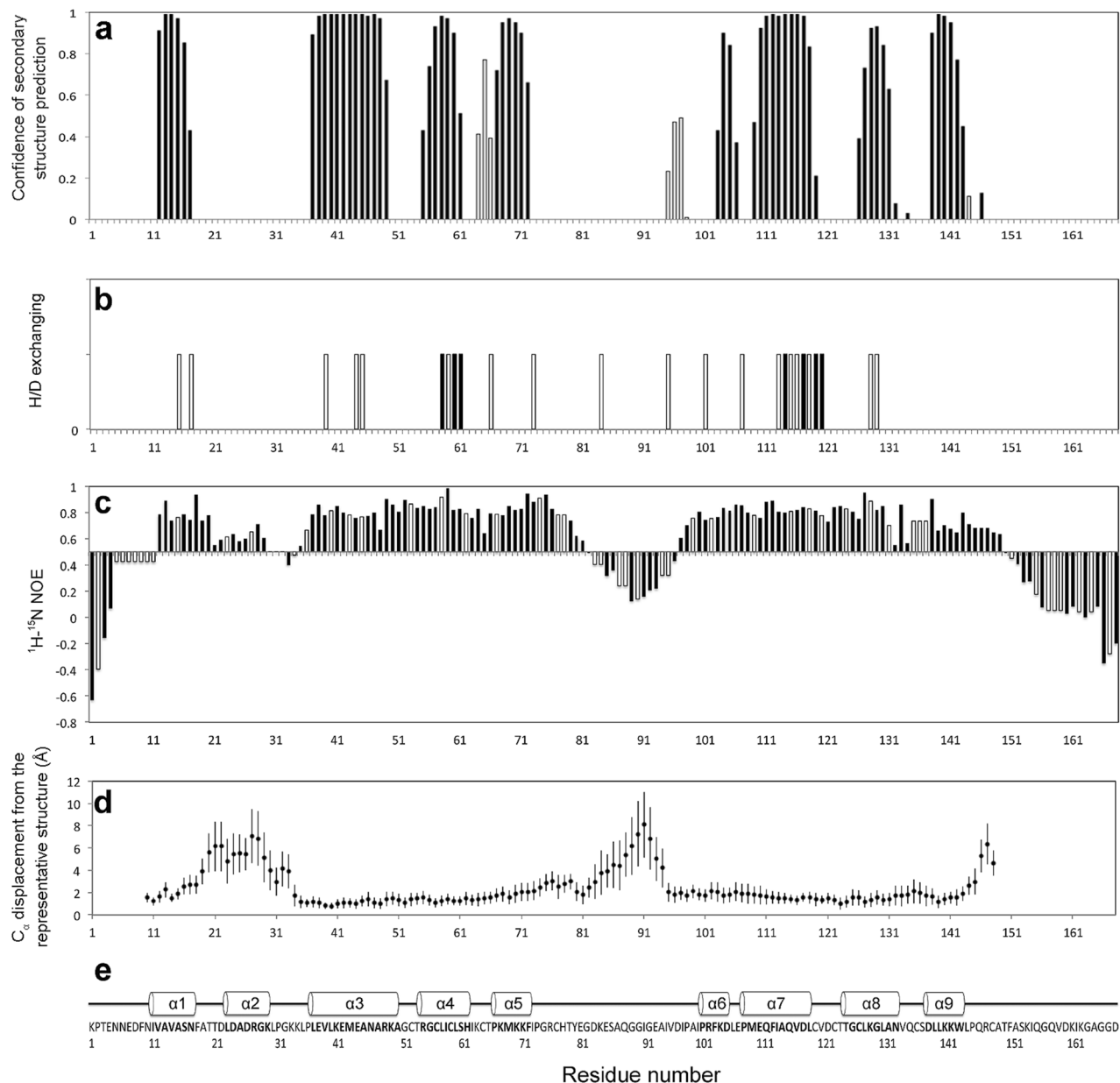
Figure 2. Residue-resolved structural and dynamics features of GLuc. (**a**) GLuc's Secondary structure predicted by TALOS+: α-helix and β-sheet tendency are shown with solid and open bars, respectively. (**b**) H/D exchange experiments. Residues that retained resonance signal after incubation in D_2O after 20 min and 18 h are marked with open bars and solid bars, respectively (^1H–^{15}N HSQC figures are shown in Supplemental Fig. 7). (**c**) ^1H–^{15}N heteronuclear NOE experiment data used to assess GLuc backbone flexibility: ^1H–^{15}N heteronuclear NOE are shown with solid bars. The NOE values of residues that were not identified were assumed using the average value of the preceding and following residues are shown with open bars. Flexible regions of GLuc were identified with the threshold value of 0.5. (**d**) Backbone C_α displacement from the representative structures is calculated using nineteen NMR-derived structures, and the error bars show standard deviations. (**e**) GLuc's amino acid sequence and secondary structure as assigned in the representative structure.

(LC–MS) analysis of the fragments. A peptide fragment with Mw. 4259.01 Da was observed and was identified as the combination of fragments A50-R54, G55-K64 and D106-K129 (Supplementary Methods, Supplementary Table 1 and Supplementary Fig. 5). This indicated presence of the C52/C127 disulfide bond, and at least another bond between C56/C123 or C59/C120 fully corroborating the NMR calculations.

Overall fold and dynamic features of GLuc. The structure with the lowest overall target function among the 20 NMR-derived structures in each round was selected. Nineteen structures from 19 rounds were used for further analysis. Among the nineteen structures, seven structures formed the putatively correct disulfide bonds (C52/C127, C56/C123, C59/C120, C65/C77, C136/C148). We finally selected the structure with the lowest aver-

NOE distance restraints*	
All	2573.4 ± 42.4
Intra residue ($|i − j| = 0$)	565.9 ± 14.7
Sequential ($|i − j| = 1$)	728.6 ± 6.6
Medium-range ($2 ≤ |i − j| ≤ 4$)	700.5 ± 18.1
Long-range ($|i − j| > 5$)	579.7 ± 32.2
Dihedral angle	183
Hydrogen bonds	25
Disulfide bonds	3
RMSD to the representative structure**	
Backbone $C_α$ (residues 10–18, 36–81, 96–145)	1.39 Å ± 0.39 Å
Heavy atoms (residues 10–18, 36–81, 96–145)	1.84 Å ± 0.20 Å
Backbone $C_α$ (helices without α2)	1.30 Å ± 0.36 Å
Heavy atoms (helices without α2)	1.74 Å ± 0.45 Å
Ramachandran plot (representative structure)	
Residues in most favored regions	71.8%
Residues in additionally allowed regions	20.4%
Residues in generously allowed regions	4.9%
Residues in disallowed regions	2.8%

Table 2. Structural statistics for the nineteen best NMR-derived GLuc structures. 19 rounds of CYANA calculation with different random seed[37]. * The averaged numbers of NOEs and their standard deviations are calculated over the 19 rounds of CYANA calculations. ** Averaged over the 18 structures (except for the representative structure).

age pairwise RMSD (against all other eighteen structures) as the representative structure. This structure also forms the putatively correct disulfide bonds.

The nineteen superimposed NMR-derived structures with the lowest overall target function show that GLuc has nine helices (α1- α9, Figs. 2e and 3), and the location of all helices essentially corroborate the TALOS + prediction except for α2 (Fig. 2a,e). The N- (residues 1–9) and C-terminus (residues 146–168) of GLuc are highly disordered (Fig. 2c). GLuc's main structure is formed by residues 10–145, in which the structure of residues 10–18, 35–81 and 97–145 were well-defined with an average backbone RMSD to the representative structure for all other eighteen structures of 1.39Å (Table 2). Residues 19–34 and 82–96 were highly disordered and can be considered as intrinsically disordered regions (IDR[27], Figs. 2d, and 3). The structures of helices α1 and α3–α9 were well-defined with an average backbone RMSD of 1.30Å (Table 2). It has been reported that α3-loop- α4-loop- α5 (α3- α5, residues 37–72) and α7-loop-α8-loop-α9 (α7–α9, residues 109–143) are repeat sequences[13,28]. The present structural analysis shows that GLuc's two repeat sequences are connected by the second IDR (residues 82–96) and form an anti-parallel bundle (α3 + α8 pair and α4 + α7 pair) that surrounds the N-terminal α1 helix. The anti-parallel bundles are firmly tied by three disulfide bonds (C52/C127, C56/C123, and C59/C120), resulting in a high local stability (Fig. 3). Moreover, all residues in the well-defined region exhibited $^1H–^{15}N$ heteronuclear NOE values larger and more uniform than residues in the N- and C- terminus or in the two IDRs, confirming that the well-defined regions obtained by calculation were consistent with the rigid regions determined by HN NOE values (Fig. 2c,d).

Discussion

The structure of GLuc is novel, as we detected no similar structures in the Protein Data Bank using DALI[29] (Supplementary Fig. 6). It is even quite different from the structures of *Renilla* luciferase (RLuc)[30], *Oplophorus* Luciferase (OLuc)[31] and apoaequorin[32], which like GLuc uses coelenterazine as a substrate and are ATP independent luciferases. The anti-parallel bundle of helices, which exhibits pseudo twofold symmetry, in the GLuc fold can be divided into two moieties. Though both showed well-defined backbone structures, the experimental data indicated differences in the side chain packing stability. The side chains of residues 52–123 are tightly packed whereas those of residues 19–51 and 124–151 are loosely packed (Fig. 3d). The high stability of the former one reveals good agreement with the residues showing extreme low H/D exchange rates, whereas residues in the latter one exhibited high H/D exchange rate indicative of a low stability (Fig. 2b and Supplementary Fig. 7). In the tightly packed moiety, we found several hydrophilic residues with well-determined side-chain structures. For instance, the chemical shifts of R76-$H_ε$, Q112-$H_ε$ 1/2, and Q116-$H_ε$ 1/2 were clearly different from averaged values observed in a flexible side chain. Furthermore, many NOEs were assigned to these atoms corroborating the fact that these side chains are involved in hydrogen bonds stabilizing the tightly packed moiety. Interestingly the side-chains of R76 and Q112 are stacked to the aromatic rings of F113 and F104, respectively, apparently shifting NMR signals of these protons from ring current effect (Fig. 3d). Finally, the hydroxyl protons of T66 and T124 were also clearly visible, suggesting that they are involved in hydrogen bonds and thus in the N-terminal capping of helices α5 and α8, respectively (Fig. 3d).
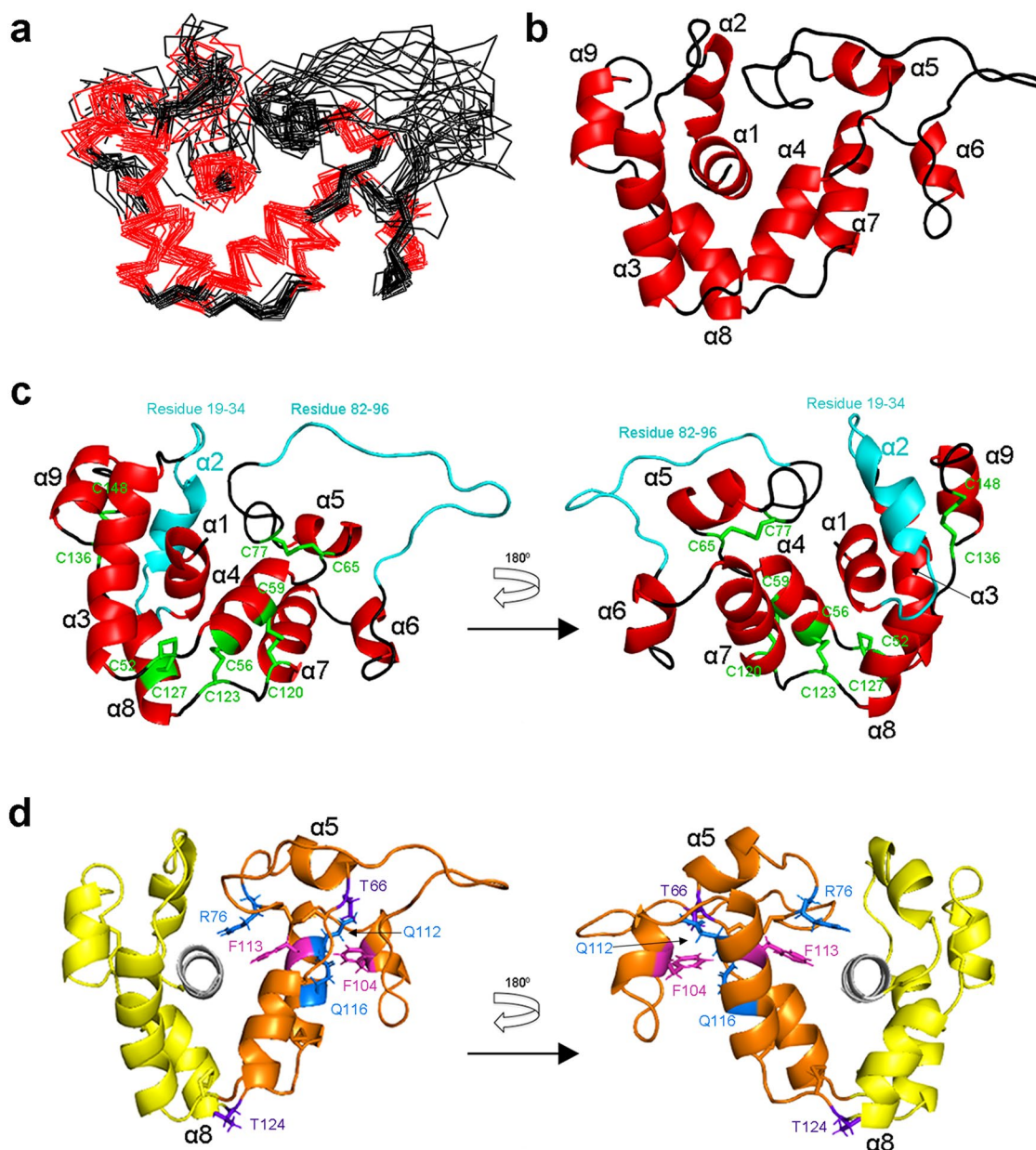
Figure 3. Overall fold of GLuc (residues 10–148) determined by NMR. The structures were generated using PyMOL. (**a**) Wire model of nineteen superimposed NMR-derived structures with the lowest target function. Helices were shown in red and loops in black. (**b**) Ribbon model of the representative structure. The nine helices of GLuc are marked from α1 to α9. (**c**) Ribbon model of the representative structure with the five disulfide bonds colored in green. Two IDRs are in cyan. (**d**) Ribbon model of the representative structure with its two moieties. The tightly packed moiety (residues 52–123) is shown in orange, whereas the loosely packed moiety (residues 19–51, 124–151) is shown in yellow. The central helix, α1 (residues 1–18), is in white. Residues R76, Q112 and Q116 are in blue, F104 and F113 are in magenta, and T66 and T124 are in purple. All structures were generated using PyMOL.

Surface accessible analysis of the representative structure indicated a noticeable cavity located among the central α1, α4 and α7 (Fig. 4a,b and Supplementary Fig. 8). The cavity was made by 19 residues: N10, V12, A13, V14, S16, N17, F18, L60, S61, I63, K64, C65, R76, C77, H78, T79, F113, I114, V117 (The hydrophobic residues are underlined for reader's convenience; Fig. 4c). Similar cavities formed by these 19 residues were identified in all other eighteen NMR-derived structures, though the sizes and shapes of their cavities showed some variation because of the limited resolution of the NMR structures. The 19 residues are distributed on three structural segments: α4 + α7 most rigid block, R76-T79 short loop, and the central α1. As mentioned above, the α4 + α7 is a rigid block stabilized by three disulfide bonds: C52/C127, C56/C123 and C59/C120 (Fig. 3c). It has been reported that GLuc N (residue 1–91) and C (residue 92–168) fragments could be used for PCAs (Protein-fragments Complementation Assays), where interactions between the fragments are detected by luminescence.
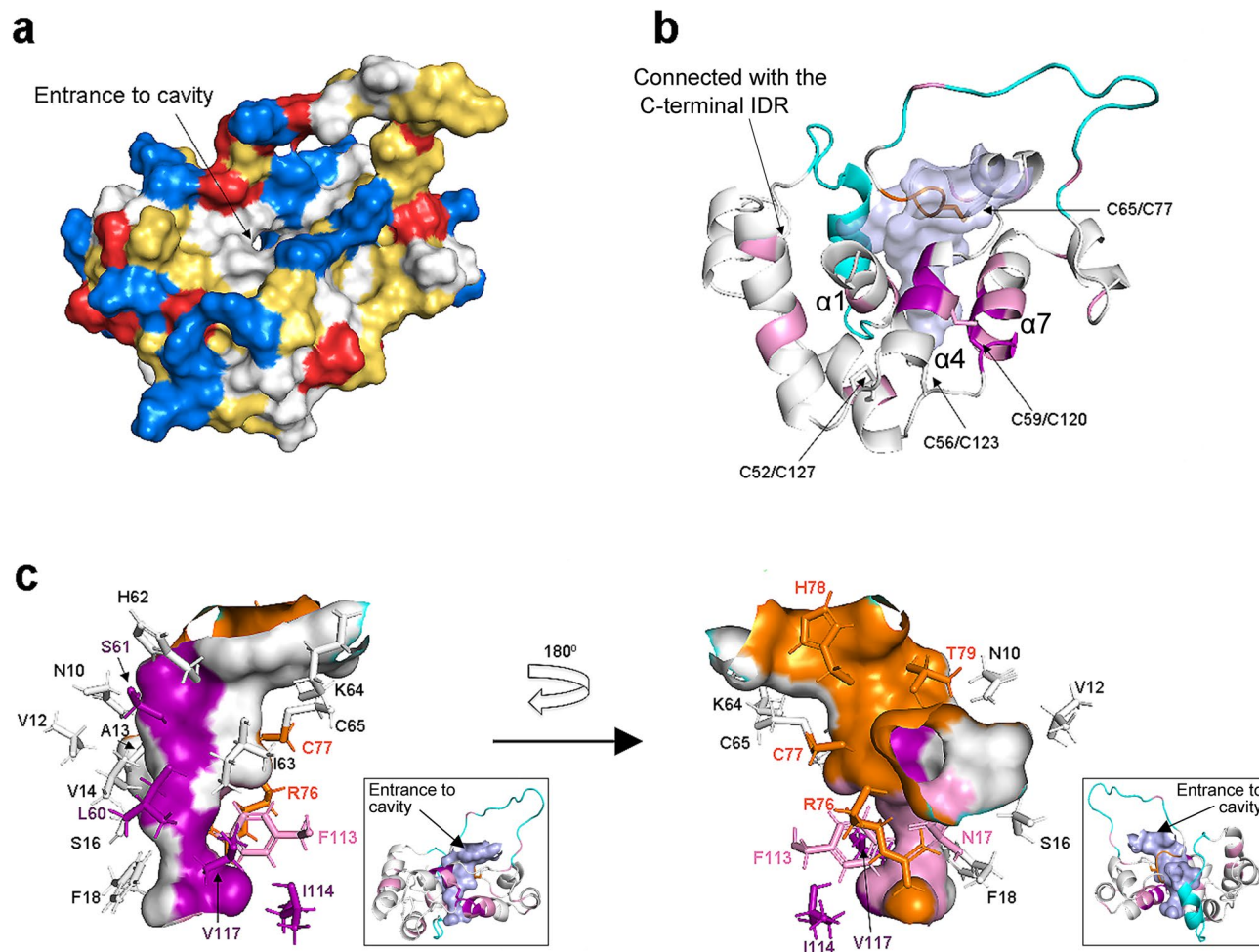
Figure 4. The cavity shown using the representative structure (residues 10–148). The structures were generated using PyMOL. (**a**) Surface representation of GLuc: positive residues (Arg, Lys and His) are colored in blue; negative residues (Glu and Asp) are colored in red; and hydrophobic residues are colored in yellow. The entrance to cavity is indicated by an arrow. (**b**) The cavity representation (colored in transparent light blue) of GLuc shown from the same direction as in (**a**). Residues that retained ^{1}H–^{15}N HSQC signals after 20 min and 18 h H/D exchanging are shown in pink and purple, respectively; residues located in the activity-related loop R76-T79 are in orange; two IDRs are shown in cyan. (**c**) Residue composition around the interior cavity. The cavity wall and its contributing residues are colored using the same color code as in (**b**). The insets show ribbon models of GLuc with the cavity colored in light blue and viewed from the same direction as in the main panel. The entrance to the cavity is indicated by an arrow. All structures were generated using PyMOL.

However, the luminescence intensity was about 10% of that of the full length, intact GLuc[33] suggesting that without the aforementioned inter-fragment disulfide bond, the cavity is formed in a very dynamic and partial manner resulting in the weaker bioluminescence. In addition to the rigid structure of the cavity, we observed a structural exchange suggested by ^{15}N relaxation dispersion (Table 1). S61 was close to both V12 and T79 (Fig. 4c), which exhibited the second-largest dispersions. We thus hypothesized that the structural exchange is related to the opened and closed form of this cavity.

A flexible docking simulation indicated that the cavity was large enough to accommodate coelenterazine; and this was verified for all seven models that formed five disulfide bonds (Supplementary Fig. 9). H/D exchange indicated that the amide protons of N17, L60, S61, F113, I114 and V117, which are located around the cavity, were visible after 20 min incubation in D$_2$O, and among them, L60, S61, I114, and V117 signals were visible even after 18hrs (Fig. 2b and Supplementary Fig. 7). The four residues are located in the α4 + α7 most rigid block (Fig. 2b,e), indicating that the cavity wall is rigid. The hydrophobic character of the cavity's interior suggests a putative role in recruiting coelenterazine, a small poorly soluble molecule. Furthermore, three activity-related residues R76, C77, H78[10] are located in the short R76-T79 loop that is near the α4 + α7 most rigid block and stabilized through the C65/C77 disulfide bond. C65 and C77 are also associated with bioluminescence activity. In particular, the mutation of C77 resulted in a vanishing luminescence[10]. This can be rationalized by hypothesizing that the breakage of the C65/C77 disulfide bond would destroy the entire cavity structure and inactivate GLuc.

Sequence alignment also suggests that the cavity plays a role as a binding pocket for coelentarzine. First, the seven cavity forming residues: L60, S61, V117 (in the hydrophobic region); R76, H78 (in the activity-related

loop); and C65, C77 (the disulfide bond) are highly conserved in 12 luciferases (MoLuc, MpLuc, etc. see Supplementary Fig. 10). C65, R76, C77, V117 were fully conserved, and L60, S61, H78 had a 92% conservation ratio. Furthermore, the structures of OLuc, RLuc and apoaequorin also contain a similar hydrophobic cavity. Altogether, these observations strongly suggest that the cavity constitutes the coelenterazine's binding site and is thus essential for GLuc's bioluminescence activity (Fig. 4c and Supplementary Fig. 10a). In addition, we noticed that the central α1 contains seven residues located near the cavity (N10, V12, A13, V14, S16, N17, F18) that are weakly conserved (Supplementary Fig. 10a), suggesting that α1 does not contribute to substrate recruiting nor to catalysis, but rather assists the cavity formation.

Additionally, we noticed that several residues in the C-terminal region (K141-D168) are remarkably conserved (Supplementary Fig. 10a) despite their high flexibility as assessed by heteronuclear NOE analysis (Fig. 2), which may suggest that they are functionally or perhaps structurally important. The sequence alignment of residues 27–97 with residues 98–168 indicates that K141-F151 in the C-terminal region has a high similarity with K70-Y80 where the aforementioned activity-related loop (R76-T79) is located (Supplementary Fig. 10b). Furthermore, our previous mutational analysis demonstrated that W143, L144 and F151 also play an important role in GLuc's activity[11], and it is of interest to note that these conserved residues are disordered in our NMR structure and can be defined as IDRs.

Finally, let us note that several lines of evidence suggested that these residues are not completely disordered. First, residues around F151 exhibited low ^1H–^{15}N NOE values ($0.0 \sim -0.2$, Fig. 2c), suggesting a disordered state in the nano- or pico-second time scale, and the R_2-dispersion experiments indicated that these residues experience a micro- or milli-second time scale exchange between the folded and unfolded states rather than in a perfectly flexible state (Table 1). This exchange between a folded and a less folded state was further corroborated by the observation of strong intra-residue and sequential NOEs in the 3D ^{15}N-edited NOESY, and the relatively broad line shapes of the peaks in the 2D ^1H–^{15}N HSQC (data not shown). Taken together, our results raise the possibility of an active participation of flexible regions (that can be considered as IDRs) in the coelenterazine oxidation reaction.

Conclusion

We produced a recombinant ^{13}C, ^{15}N labeled GLuc in *E.coli*, and assigned nearly all of the backbone and most of the side chain chemical shifts. The N- and C-termini, as well as the segment located between α1 and α3 (encompassing α2) and the loop between α5 and α6 were flexible. GLuc's structure is unique and is made of nine helices, constituting two anti-parallel bundles, which are formed by, respectively, helices α3-α4 and α7-α8 of parts of homologous sequential repeats. The helices are tied together by disulfide bonds to form a 4-finger structure with a pseudo-twofold symmetry surrounding the N-terminal helix 1. Finally, we identified a hydrophobic cavity where coelenterazine is most likely to bind and the catalytic reaction occurs, but it is difficult to discuss whether one or two coelenterazines can bind to the cavity, and how the isolated N-and C-terminal regions can retain some activity[28,34]. Presently, our structure does not provide preference for a single or a multiple binding site model. In addition, the large and fluctuating nature (Table 1) of the cavity and because the activity-related C-terminal IDR is located somewhat far from the cavity (Fig. 4b) may result in a catalytic reaction that does not adhere to traditional binding models. The fold of GLuc is novel, and we believe that the above reported structural/dynamic information will open an avenue for redesigning the bioluminescence activity of GLuc and thereby widen its scope of application.

Methods

Expression system. A DNA sequence encoding the wild-type GLuc gene (UniProtKB ID: Q9BLZ2) without the 17 residues secretion tag and with an E100A and G103R mutations that increased protein expression was synthesized as reported previously[13]. The GLuc sequence was flanked with an N terminal His-tag and a C terminal SEP-tag (Solubility Enhancement Peptide tag, C9D) to facilitate protein expression, refolding, and purification[20]. Two Factor Xa cleavage sites were inserted between GLuc and His-tag/SEP-Tag. The GLuc gene named GLuc-TG[13] was inserted into pET21c (Novagen) at the NdeI/BamHI site to construct p21GLucTG with ampicillin resistance.

Protein expression and purification. p21GLucTG was transformed into BL21(DE3), and pre-cultured in 1 L Luria–Bertani (LB) medium at 37 °C and 250 rpm shaking. When OD_{590nm} reached 1.0, *E. coli* cells were collected by soft centrifugation and transferred to a 1 L M9 medium containing ^{13}C-glucose and $^{15}NH_4Cl$. Isopropyl β-D-Thiogalactoside (IPTG) was added at 1 mM final concentration for inducing protein expression, and the temperature was lowered to 25 °C for minimizing the formation of inclusion bodies. After 4 h with shaking at 250 rpm, the cells were harvested by centrifugation and sonicated. GLuc was purified from the supernatant fraction using a Nickel Nitrilotriacetic Acid (NTA) column followed with overnight dialysis at 4 °C against 50 mM Tris–HCl, pH 8.0. GLuc was then air-oxidized for 3 days at the same conditions in order to form the five disulfide bridges. Residual misfolded GLuc was removed using a reversed phase High-Performance Liquid Chromatography (HPLC). The protein concentration was determined using a Bradford assay[35], and Factor Xa was added to GLuc dissolved in 50 mM Tris–HCl, 100 mM NaCl, and 5 mM $CaCl_2$ at a ratio of 1:100 (w/w), and the sample was again incubated for 8 h at 37 °C, 100 rpm for enzymatic cleavage of the His- and the SEP-Tags. Uncleaved GLuc was removed using, again, reversed phase HPLC. GLuc identity was confirmed by MALDI-TOF mass spectroscopy on an ABI SCIEX TOF/TOF 5800 (Thermo Fisher Scientific Inc., Massachusetts, USA). GLuc folding was confirmed by bioluminescence activity measurement on a FP-8000 fluorescence spectrophotometer (JASCO International co., Ltd, Tokyo, Japan). GLuc was freeze-dried and kept as a powder at − 30 °C until use.

Analytical Ultracentrifugation (AUC).   Sedimentation equilibrium experiments were carried out using an Optima XL-A analytical ultracentrifuge (Beckman-Coulter, Inc., Brea, California, USA) with a four-hole An60Ti rotor at 20 °C. Before centrifugation, GLuc samples were dialyzed overnight against 50 mM MES and 100 mM NaCl at pH 4.7. The solvent density (1.006983 g/cm^3) was determined using DMA 5000 (Anton Paar). Each sample was then transferred into a cell with a six-channel centerpiece. The sample concentrations were 1.2, 0.6, and 0.3 mg/mL. Data were obtained at 12,000, 22,000, and 37,000 rpm. A total equilibration time of 24 h was used for each speed, with absorbance scans at 280 nm taken every 4 h to ensure that equilibrium had been reached. Data analysis was performed by global analysis of all of the data sets obtained at different concentrations and rotor speeds using SEDPHAT[36].

NMR analysis and structure calculation.   NMR experiments for resonance assignments and ^{1}H–^{15}N heteronuclear NOE experiment were conducted using 0.2 mM ^{15}N single or ^{15}N, ^{13}C double labeled GLuc protein dissolved in 50 mM MES buffer pH 6.0 and 2 mM NaN$_3$, at 293 K with 8%(v/v) D$_2$O in a 5 mm Shigemi microtube (Shigemi co., Ltd, Tokyo, Japan). NMR spectra were acquired on a Bruker Avance-III 700 MHz spectrometer, equipped with a 5 mm CPTXI cryoprobe. Two-dimensional and three-dimensional NMR experiments (^{1}H–^{15}N HSQC, HNCACB, CBCA(CO)NH, HNCA, HNCO, HN(CA)CO) were performed for the backbone ^{15}N and ^{13}C assignments. ^{15}N–TOCSY–HSQC, ^{15}N–NOESY–HSQC, and HCCH–TOCSY were used for backbone and side-chain signal assignments. H/D exchange ^{1}H–^{15}N–HSQC experiment was performed under the above-described conditions but by dissolving GLuc's freeze-dried powder in D$_2$O instead of H$_2$O. The transverse relaxation rate (R_2) dispersion experiments for backbone ^{15}N atoms were performed on a Bruker Avance-III 900 MHz spectrometer, using pulse scheme including constant time relaxation compensated CPMG pulse sequences[37]. A series of 2D experiments were acquired with various rates of CPMG pulse (50, 100, 150, 200, 250, 300, 400, 500, 600, 800, 1000 Hz) in the two of 20 ms CPMG blocks. The ^{15}N CPMG irradiation intensity was set to 3125 Hz (~35 ppm). To avoid the error in peak intensity arisen from the offset effect, two sets of relaxation experiments with different ^{15}N irradiation centers (111 ppm and 125 ppm) were carried out. For each signal, we read the series of peak intensities using the spectrum giving the smaller errors by the offset effect.

Three-dimensional structure determination.   Automated NOE assignments and structure calculations were performed using CYANA (ver. 3.98) on a PC-cluster equipped with 20-core Intel Xeon E5-4627v3 (3.0 GHz) and using the manually assigned chemical shifts and a list of NOE chemical shifts derived from the 3D ^{15}N-, ^{13}C-edited NOESY spectra of the aliphatic and aromatic regions. For each cycle of CYANA calculation, 20 out of 100 structures were selected after 10,000 steps of simulated annealing using distance constraints derived from automatically assigned NOEs. The detailed algorithm and strategy are described[23]. For the automated NOE assignments of the NOEs peaks, the tolerances were set to 0.04, 0.4 and 0.4 ppm for ^{1}H, ^{15}N and ^{13}C signals, respectively. 19 rounds of CYANA calculations with different random seeds were performed. It should be noted that, for a minor number of the NOEs, the automated assignment was ambiguous and depended on the random seed. We thus calculated the structures using restraints that were slightly different from set to set, and we selected the best structure from the structures generated using each of the sets and reported the ensemble of structures. The treatment using many random seeds was previously discussed and the ensemble of structures calculated using ambiguous assignments becomes more diverse and safer. We can reduce the risk to make structure ensemble affected by wrong NOE assignments[38]. According to experimentally determined slow-exchanging backbone amide protons and threonine (Thr) side chain OH atoms, we applied 24 sets of distance constraints, two of them for hydrogen bonds including Thr-OH related to N-terminal capping of α-helices and 21 of them for backbone amide protons to the 19 rounds of CYANA calculations.

Other softwares.   The secondary structure elements of GLuc were predicted by submitting the backbone chemical shifts into TALOS+ (https://spin.niddk.nih.gov/bax/nmrserver/talos/)[39]. Root-Mean-Square Deviation (RMSD) was calculated using a Biopython.PDB module (https://biopython.org)[40]. Three-dimensional images of the NMR structures were generated using PyMOL. The flexible docking simulation was calculated using AutoDock (Ver. 4.2.6) and AutoDockTools (Ver.1.5.6)[41].

References
1. Shimomura, O. *Bioluminescence: Chemical Principles and Methods*, vol (World Scientific, Singapore, 2006).
2. Oyama, H. *et al. Gaussia* luciferase as a genetic fusion partner with antibody fragments for sensitive immunoassay monitoring of clinical biomarkers. *Anal. Chem.* **87**, 12387–12395 (2015).
3. Kaskova, Z. M., Tsarkova, A. S. & Yampolsky, I. V. 1001 lights: luciferins, luciferases, their mechanisms of action and applications in chemical analysis, biology and medicine. *Chem. Soc. Rev.* **45**, 6048–6077 (2016).
4. Yi, S., Liu, N.-N., Hu, L., Wang, H. & Sahni, N. Base-resolution stratification of cancer mutations using functional variomics. *Nat. Protoc.* **12**, 2323 (2017).
5. Sent-Gyorgyi, C. S. & Luciferases, B. J. B. fluorescent proteins, nucleic acids encoding the luciferases and fluorescent proteins and the use thereof in diagnostics, high throughput screening and novelty items. U.S. Patent 6232107-B (2001).
6. Tannous, B. A., Kim, D. E., Fernandez, J. L., Weissleder, R. & Breakefield, X. O. Codon-optimized *Gaussia* luciferase cDNA for mammalian gene expression in culture and in vivo. *Mol. Ther.* **11**, 435–443 (2005).

7. Maguire, C. A. *et al.* *Gaussia* luciferase variant for high-throughput functional screening applications. *Anal. Chem.* **81**, 7102–7106 (2009).

8. Welsh, J. P., Patel, K. G., Manthiram, K. & Swartz, J. R. Multiply mutated *Gaussia* luciferases provide prolonged and intense bio-luminescence. *Biochem. Biophys. Res. Commun.* **389**, 563–568 (2009).

9. Degeling, M. H. *et al.* Directed molecular evolution reveals *Gaussia* luciferase variants with enhanced light output stability. *Anal. Chem.* **85**, 3006–3012 (2013).

10. Kim, S. B., Suzuki, H., Sato, M. & Tao, H. Superluminescent variants of marine luciferases for bioassays. *Anal. Chem.* **83**, 8732–8740 (2011).

11. Wu, N., Kamioka, T. & Kuroda, Y. A novel screening system based on VanX-mediated autolysis-Application to *Gaussia* luciferase. *Biotechnol. Bioeng.* **113**, 1413–1420 (2016).

12. Gheysens, O. & Mottaghy, F. M. Method of bioluminescence imaging for molecular imaging of physiological and pathological processes. *Methods* **48**, 139–145 (2009).

13. Wu, N., Rathnayaka, T. & Kuroda, Y. Bacterial expression and re-engineering of *Gaussia princeps* luciferase and its use as a reporter protein. *BBA-Proteins Proteom.* **1854**, 1392–1399 (2015).

14. Tannous, B. A. *Gaussia* luciferase reporter assay for monitoring biological processes in culture and in vivo. *Nat. Protoc.* **4**, 582–591 (2009).

15. Goerke, A. R., Loening, A. M., Gambhir, S. S. & Swartz, J. R. Cell-free metabolic engineering promotes high-level production of bioactive *Gaussia princeps* luciferase. *Metab. Eng.* **10**, 187–200 (2008).

16. Rathnayaka, T., Tawa, M., Sohya, S., Yohda, M. & Kuroda, Y. Biophysical characterization of highly active recombinant *Gaussia* luciferase expressed in Escherichia coli. *BBA-Proteins Proteom.* **1804**, 1902–1907 (2010).

17. Islam, M. M., Khan, M. A. & Kuroda, Y. Analysis of amino acid contributions to protein solubility using short peptide tags fused to a simplified BPTI variant. *BBA-Proteins Proteom.* **1824**, 1144–1150 (2012).

18. Khan, M. A., Islam, M. M. & Kuroda, Y. Analysis of protein aggregation kinetics using short amino acid peptide tags. *BBA-Proteins Proteom.* **1834**, 2107–2115 (2013).

19. Nautiyal, K. & Kuroda, Y. A SEP tag enhances the expression, solubility and yield of recombinant TEV protease without altering its activity. *New Biotechnol.* **42**, 77–84 (2018).

20. Rathnayaka, T. *et al.* Solubilization and folding of a fully active recombinant *Gaussia* luciferase with native disulfide bonds by using a SEP-Tag. *BBA-Proteins Proteom.* **1814**, 1775–1778 (2011).

21. Bah, A. *et al.* Folding of an intrinsically disordered protein by phosphorylation as a regulatory switch. *Nature* **519**, 106 (2014).

22. Yi, Q. & Baker, D. Direct evidence for a two-state protein unfolding transition from hydrogen-deuterium exchange, mass spectrometry, and NMR. *Protein Sci.* **5**, 1060–1066 (2008).

23. Güntert, P. & Buchner, L. Combined automated NOE assignment and structure calculation with CYANA. *J. Biomol. NMR* **62**, 453–471 (2015).

24. Schmidt, E. & Güntert, P. A new algorithm for reliable and general NMR resonance assignment. *J. Am. Chem. Soc.* **134**, 12817–12829 (2012).

25. Johnson, B. A. & Blevins, R. A. NMR View: a computer program for the visualization and analysis of NMR data. *J. Biomol. NMR* **4**, 603–614 (1994).

26. Kobayashi, N. *et al.* KUJIRA, a package of integrated modules for systematic and interactive analysis of NMR data directed to high-throughput NMR structure studies. *J. Biomol. NMR* **39**, 31–52 (2007).

27. Ota, M. *et al.* An assignment of intrinsically disordered regions of proteins based on NMR structures. *J. Struct. Biol.* **181**, 29–36 (2013).

28. Inouye, S. & Sahara, Y. Identification of two catalytic domains in a luciferase secreted by the copepod *Gaussia princeps*. *Biochem. Biophys. Res. Commun.* **365**, 96–101 (2008).

29. Hasegawa, H. & Holm, L. Advances and pitfalls of protein structural alignment. *Curr. Opin. Struc. Biol.* **19**, 341–348 (2009).

30. Loening, A. M., Fenn, T. D. & Gambhir, S. S. Crystal structures of the luciferase and green fluorescent protein from *Renilla reniformis*. *J. Mol. Biol.* **374**, 1017–1028 (2007).

31. Tomabechi, Y. *et al.* Crystal structure of nanoKAZ: the mutated 19 kDa component of *Oplophorus* luciferase catalyzing the bioluminescent reaction with coelenterazine. *Biochem. Biophys. Res. Commun.* **470**, 88–93 (2016).

32. Head, J. F., Inouye, S., Teranishi, K. & Shimomura, O. The crystal structure of the photoprotein aequorin at 2.3 Å resolution. *Nature* **405**, 372–376 (2000).

33. Remy, I. & Michnick, S. W. A highly sensitive protein-protein interaction assay based on *Gaussia* luciferase. *Nat. Methods* **3**, 977–979 (2006).

34. Hunt, E. A. *et al.* Truncated variants of *Gaussia* luciferase with tyrosine linker for site-specific bioconjugate applications. *Sci. Rep.* **6**, 26814 (2016).

35. Bradford, M. M. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal. Biochem.* **72**, 248–254 (1976).

36. Vistica, J. *et al.* Sedimentation equilibrium analysis of protein interactions with global implicit mass conservation constraints and systematic noise decomposition. *Anal. Biochem.* **326**, 234–256 (2004).

37. Tollinger, M., Skrynnikov, N. R., Mulder, F. A. A., Forman-Kay, J. D. & Kay, L. E. Slow dynamics in folded and unfolded states of an SH3 domain. *J. Am. Chem. Soc.* **123**, 11341–11352 (2001).

38. Buchner, L. & Güntert, P. Increased reliability of nuclear magnetic resonance protein structures by consensus structure bundles. *Structure* **23**, 425–434 (2015).

39. Shen, Y., Delaglio, F., Cornilescu, G. & Bax, A. TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. *J. Biomol. NMR* **44**, 213–223 (2009).

40. Manderick, B. & Hamelryck, T. PDB file parser and structure class implemented in Python. *Bioinformatics* **19**, 2308–2310 (2003).

41. Morris, G. M. *et al.* AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. *J. Comput. Chem.* **30**, 2785–2791 (2009).

Acknowledgements

Author contributions

N.W., Y.K. and T.Y. conceived the project, analyzed the structure, and wrote the manuscript. N.W., K.T., and T.Y. performed the NMR experiments and analyzed the data, N.K and T.Y. performed and analyzed the structure calculation with N.W.'s assistance, S.U. and T.S. performed the AUC experiments. T.Y. and T.S. performed the

LC-MS analysis. All authors contributed to finalize the manuscript and approved it. *Data deposition* The chemical shifts have been deposited in the Biological Magnetic Resonance Bank (BMRB) under the accession No.36288, and the atomic coordinates are deposited in the Protein Data Bank under accession number PDB-ID: 7D2O. The expression vector for GLuc-TG (p21GLucTG) is deposited in Addgene (ID:124660).

Funding

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-76486-4.

Correspondence and requests for materials should be addressed to Y.K. or T.Y.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.