expression of diabetes-related genes, such as leptin (22), the adiponectin receptor (23), PPAR (24), and plasminogen activator inhibitor-1 (25), was observed to follow circadian patterns. Likewise, we have demonstrated for the first time that Angptl2 is a novel circadian gene involved in the pathology of type 2 diabetes mellitus. On the basis of these reports, we hypothesized that nutrient homeostasis may be maintained by a daily cycle of increases and decreases in the expression of metabolic regulating factors. We therefore examined the effects of Angptl2 on diabetes *in vivo* by administration of the protein, rather than using transgenic or knockout mouse models.

In 3T3-L1 adipocytes, Angptl2 induced insulin sensitivity of Akt. Although the mechanism of this action is not entirely clear, Trib3 expression is known to be increased by treatment with Angptl2 siRNA. There is also evidence that Trib3 siRNA partially recovers Angptl2 siRNA-induced insulin resistance in 3T3-L1 adipocytes. In addition, Trib3 has also been shown to interfere with insulin signaling by binding to and inhibiting Akt (22, 26). Trib3 expression can be increased by numerous stimuli, including starvation (27), PPAR-mediated activation (26), and chronic alcohol ingestion (28), all of which are linked to insulin insensitivity. Overexpression of Trib3 in a cell line moderately decreased insulin-induced phosphorylation and activation of Akt (20). Furthermore, it has been reported that Trib3 suppresses adipocyte differentiation by negatively regulating PPARγ transcriptional activity (29). Taken together, these findings indicate that the regulation of Trib3 expression may be one of the mechanisms by which Angptl2 improves insulin resistance. However, long-term treatment with Angptl2 did not stimulate phosphor-Akt, whereas chronic treatment with Angptl2 did acutely stimulate phosphor-Akt (Supplemental Fig. 7). Furthermore, acute treatment with Angptl2-induced Akt phosphorylation was suppressed by pretreatment with LY294002, a phosphatidylinositol 3 kinase inhibitor (Supplemental Fig. 7). These results suggest that different mechanisms exist between acute and continuous treatment-induced Akt phosphorylation. Thus, further studies are needed to investigate the mechanisms involved in Angptl2-stimulated insulin signaling.

In contrast to our findings, a recent study showed that Angptl2 promotes chronic adipose tissue inflammation and obesity-related systemic insulin resistance (30). In this article, Angptl2-transgenic mice showed inflammation in adipose tissue and insulin resistance. Conversely, Angptl2 knockout mice showed reduced inflammation after high-fat feeding. Angptl2 activates chemotaxis of monocytes/macrophages result in inflammation (30). Although the results of their and our studies seem to be discrepant, we suggest that the transgenic and knockout mice used in their study represent a distinct disease state from that in db/db mice. Angptl2 transgenic and knockout mice develop diabetes or a diabetic-resistant state as they grow, whereas db/db mice show severe diabetes and inflammatory state by 8 wk of age, before the administration of Angptl2 in our study. Therefore, Angptl2 transgenic and knockout mice are useful to investigate the potential role of Angptl2 in the pathogenesis of diabetes, whereas the administration of Angptl2 in db/db mice helps us to examine the role of Angptl2 in the progression of diabetes. Here, we demonstrated that the administration of Angptl2 improved blood glucose levels, lipid metabolism, and insulin resistance in db/db mice. In cultured adipocytes, as in db/db mice, treatment with Angptl2 improved insulin sensitivity. The *db/db* mice show severe inflammation, including the activation of chemotaxis. Thus, the administration of Angptl2 may not further accelerate monocyte/macrophage chemotaxis. In fact, the infiltration of macrophages into adipose tissues and CD68 expression were not affected by Angptl2 administration. On the other hand, Angptl2 did decrease plasma lipid and free fatty acid levels and lowered the expression of cytokines and chemokines. Taken together, these results suggest that Angptl2 improves the inflammatory state and insulin sensitivity in *db/db* mice. Taken together, our results suggested that Angptl2 plays differing roles in different stages in the pathogenesis of type 2 diabetes.

In summary, we demonstrated that Angptl2 increases both adiponectin expression and insulin sensitivity, thereby reducing blood glucose levels and lipid content. These results suggest that Angptl2 plays a crucial role in regulating insulin sensitivity and is, therefore, involved in the progression of type 2 diabetes mellitus. Characterization of the Angptl2 receptor and the cellular mechanisms underlying Angptl2-induced modulation of metabolic gene expression may enable us to develop new therapies for metabolic diseases.

Acknowledgments

We thank Tomoko Kojima for her technical assistance.

Address all correspondence and requests for reprints to: Masashi Kitazawa, 2-4-7 Aomi, Koto-ku, Tokyo 135-0064, Japan. E-mail: masashi-kitazawa@aist.go.jp.

This study was performed as a part of a research and development project of the Industrial Science and Technology Program supported by the New Energy and Industrial Technology Development Organization.

Disclosure Summary: The authors have nothing to disclose.

References

1. Koishi R, Ando Y, Ono M, Shimamura M, Yasumo H, Fujiwara T, Horikoshi H, Furukawa H 2002 Angptl3 regulates lipid metabolism in mice. Nat Genet 30:151–157
2. Yoshida K, Shimizugawa T, Ono M, Furukawa H 2002 Angiopoietin-like protein 4 is a potent hyperlipidemia-inducing factor in mice and inhibitor of lipoprotein lipase. J Lipid Res 43:1770–1772
3. Oike Y, Akao M, Yasunaga K, Yamauchi T, Morisada T, Ito Y, Urano T, Kimura Y, Kubota Y, Maekawa H, Miyamoto T, Miyata K, Matsumoto S, Sakai J, Nakagata N, Takeya M, Koseki H, Ogawa Y, Kadowaki T, Suda T 2005 Angiopoietin-related growth factor antagonizes obesity and insulin resistance. Nat Med 11:400–408
4. Kitazawa M, Ohizumi Y, Oike Y, Hishinuma T, Hashimoto S 2007 Angiopoietin-related growth factor suppresses gluconeogenesis through the Akt/forkhead box class O1-dependent pathway in hepatocytes. J Pharmacol Exp Ther 323:787–793
5. Seaman GV, Engel R, Swank RL, Hissen W 1965 Circadian periodicity in some physicochemical parameters of circulating blood. Nature 207:833–835
6. Panda S, Antoch MP, Miller BH, Su AI, Schook AB, Straume M, Schultz PG, Kay SA, Takahashi JS, Hogenesch JB 2002 Coordinated transcription of key pathways in the mouse by the circadian clock. Cell 109:307–320
7. Albrecht U, Eichele G 2003 The mammalian circadian clock. Curr Opin Genet Dev 13:271–277
8. Zvonic S, Ptitsyn AA, Conrad SA, Scott LK, Floyd ZE, Kilroy G, Wu X, Goh BC, Mynatt RL, Gimble JM 2006 Characterization of peripheral circadian clocks in adipose tissues. Diabetes 55:962–970
9. Ueda HR, Hayashi S, Chen W, Sano M, Machida M, Shigeyoshi Y, Iino M, Hashimoto S 2005 System-level identification of transcriptional circuits underlying mammalian circadian clocks. Nat Genet 37:187–192
10. Turek FW, Joshu C, Kohsaka A, Lin E, Ivanova G, McDearmon E, Laposky A, Losee-Olson S, Easton A, Jensen DR, Eckel RH, Takahashi JS, Bass J 2005 Obesity and metabolic syndrome in circadian Clock mutant mice. Science 13:1043–1045
11. Rudic RD, McNamara P, Curtis AM, Boston RC, Panda S, Hogenesch JB, Fitzgerald GA 2004 BMAL1 and CLOCK, two essential components of the circadian clock, are involved in glucose homeostasis. PLoS Biol 2:e377
12. Shimba S, Ishii N, Ohta Y, Ohno T, Watabe Y, Hayashi M, Wada T, Aoyagi T, Tezuka M 2005 Brain and muscle Arnt-like protein-1 (BMAL1), a component of the molecular clock, regulates adipogenesis. Proc Natl Acad Sci USA 102:12071–12076
13. Kohsaka A, Laposky AD, Ramsey KM, Estrada C, Joshu C, Kobayashi Y, Turek FW, Bass J 2007 High-fat diet disrupts behavioral and molecular circadian rhythms in mice. Cell Metab 6:414–421
14. Kim I, Moon SO, Koh KN, Kim H, Uhm CS, Kwak HJ, Kim NG, Koh GY 1999 Molecular cloning, expression, and characterization of angiopoietin-related protein. angiopoietin-related protein induces endothelial cell sprouting. J Biol Chem 274:26523–26528
15. Kubota Y, Oike Y, Satoh S, Tabata Y, Niikura Y, Morisada T, Akao M, Urano T, Ito Y, Miyamoto T, Nagai N, Koh GY, Watanabe S, Suda T 2005 Cooperative interaction of Angiopoietin-like proteins 1 and 2 in zebra fish vascular development. Proc Natl Acad Sci USA 102:13502–13507
16. Hato T, Tabata M, Oike Y 2008 The role of angiopoietin-like proteins in angiogenesis and metabolism. Trends Cardiovasc Med 18: 6–14
17. Watson RT, Pessin JE 2006 Bridging the GAP between insulin signaling and GLUT4 translocation. Trends Biochem Sci 31:215–222
18. Dugani CB, Klip A 2005 Glucose transporter 4: cycling, compartments and controversies. EMBO Rep 6:1137–1142
19. Brunet A, Bonni A, Zigmond MJ, Lin MZ, Juo P, Hu LS, Anderson MJ, Arden KC, Blenis J, Greenberg ME 1999 Akt promotes cell survival by phosphorylating and inhibiting a Forkhead transcription factor. Cell 96:857–868
20. Du K, Herzig S, Kulkarni RN, Montminy M 2003 TRB3: a tribbles homolog that inhibits Akt/PKB activation by insulin in liver. Science 300:1574–1577
21. Kennaway DJ, Owens JA, Voultsios A, Boden MJ, Varcoe TJ 2007 Metabolic homeostasis in mice with disrupted Clock gene expression in peripheral tissues. Am J Physiol Regul Integr Comp Physiol 293:R1528–R1537
22. Ahima RS, Prabakaran D, Flier JS 1998 Postnatal leptin surge and regulation of circadian rhythm of leptin by feeding. Implications for energy homeostasis and neuroendocrine function. J Clin Invest 101: 1020–1027
23. Blüher M, Fasshauer M, Kralisch S, Schön MR, Krohn K, Paschke R 2005 Regulation of adiponectin receptor R1 and R2 gene expression in adipocytes of C57BL/6 mice. Biochem Biophys Res Commun 329:1127–1132
24. Yang X, Downes M, Yu RT, Bookout AL, He W, Straume M, Mangelsdorf DJ, Evans RM 2006 Nuclear receptor expression links the circadian clock to metabolism. Cell 126:801–810
25. Maemura K, de la Monte SM, Chin MT, Layne MD, Hsieh CM, Yet SF, Perrella MA, Lee ME 2000 CLIF, a novel cycle-like factor, regulates the circadian oscillation of plasminogen activator inhibitor-1 gene expression. J Biol Chem 275:36847–36851
26. Koo SH, Satoh H, Herzig S, Lee CH, Hedrick S, Kulkarni R, Evans RM, Olefsky J, Montminy M 2004 PGC-1 promotes insulin resistance in liver through PPAR-α-dependent induction of TRB-3. Nat Med 10:530–534
27. Fleming I 2008 Double tribble: two TRIB3 variants, insulin, Akt, and eNOS. Arterioscler Thromb Vasc Biol 28:1216–1218
28. He L, Simmen FA, Mehendale HM, Ronis MJ, Badger TM 2006 Chronic ethanol intake impairs insulin signaling in rats by disrupting Akt association with the cell membrane: role of TRB3 in inhibition of Akt/protein kinase B activation. J Biol Chem 281:11126–11134
29. Takahashi Y, Ohoka N, Hayashi H, Sato R 2008 TRB3 suppresses adipocyte differentiation by negatively regulating PPARγ transcriptional activity. J Lipid Res 49:880–892
30. Tabata M, Kadomatsu T, Fukuhara S, Miyata K, Ito Y, Endo M, Urano T, Zhu HJ, Tsukano H, Tazume H, Kaikita K, Miyashita K, Iwawaki T, Shimabukuro M, Sakaguchi K, Ito T, Nakagata N, Yamada T, Katagiri H, Kasuga M, Ando Y, Ogawa H, Mochizuki N, Itoh H, Suda T, Oike Y 2009 Angiopoietin-like protein 2 promotes chronic adipose tissue inflammation and obesity-related systemic insulin resistance. Cell Metab 10:178–188

Histone chaperone Spt6 is required for class switch recombination but not somatic hypermutation

Il-mi Okazaki[a,1], Katsuya Okawa[b,2], Maki Kobayashi[a], Kiyotsugu Yoshikawa[a,3], Shimpei Kawamoto[a,4], Hitoshi Nagaoka[a], Reiko Shinkura[a,5], Yoko Kitawaki[a], Hisaaki Taniguchi[c], Tohru Natsume[d], Shun-Ichiro Iemura[d], and Tasuku Honjo[a,6]

[a]Department of Immunology and Genomic Medicine and [b]Biomolecular Characterization Unit, Frontier Technology Center, Horizontal Medical Research Organization, Graduate School of Medicine, Kyoto University, Kyoto 606-8501, Japan; [c]Division of Disease Proteomics, Institute for Enzyme Research, University of Tokushima, 770-8503, Japan; and [d]Biomedical Information Research Center, National Institute of Advanced Industrial Science and Technology, Tokyo 135-0064, Japan

Contributed by Tasuku Honjo, April 5, 2011 (sent for review March 4, 2011)

Activation-induced cytidine deaminase (AID) is shown to be essential and sufficient to induce two genetic alterations in the Ig loci: class switch recombination (CSR) and somatic hypermutation (SHM). However, it is still unknown how a single-molecule AID differentially regulates CSR and SHM. Here we identified Spt6 as an AID-interacting protein by yeast two-hybrid screening and immunoprecipitation followed by mass spectrometry. Knockdown of Spt6 resulted in severe reduction of CSR in both the endogenous Ig locus in B cells and an artificial substrate in fibroblast cells. Conversely, knockdown of Spt6 did not reduce but slightly enhanced SHM in an artificial substrate in B cells, indicating that Spt6 is required for AID to induce CSR but not SHM. These results suggest that Spt6 is involved in differential regulation of CSR and SHM by AID.

The Ig genes in antigen-stimulated B lymphocytes are diversified by two distinct genetic alteration mechanisms, namely somatic hypermutation (SHM) and class switch recombination (CSR) (1, 2). SHM causes the accumulation of point mutations in the rearranged variable (V) region genes, leading to generation of antibodies with higher affinity after cellular selection by a limited amount of antigen (1). CSR replaces the heavy chain constant region (C_H) gene proximal to the V_H gene, namely Cμ with one of the downstream C_H genes by recombination between the switch (S) regions located 5′ to each C_H gene, thereby producing antibodies with diverse effector functions without changing their antigen specificity (2).

Both SHM and CSR require activation-induced cytidine deaminase (AID), which is specifically expressed in activated B cells (3). It is well accepted that AID initiates single-strand DNA breaks essential for SHM and CSR through its cytidine deaminase activity. AID can also introduce mutations in such non-Ig loci as *c-myc*, *Pim1*, *Pax5*, *bcl-6*, and *RhoH* (4, 5). The number of target loci of AID seems to be larger than expected but still limited (5). Although extensive analyses have been done to uncover the exact molecular mechanism how AID induces DNA strand breaks at restricted loci, it is unknown how a single-molecule AID differentially regulates CSR and SHM or how the Ig genes and other target loci are preferentially targeted in the whole genome. To answer these questions, extensive studies were carried out to identify cofactor(s) that may account for the target specificity of the AID function. Several AID-interacting proteins have been reported, including RNA polymerase II (6), replication protein A (7), protein kinase A (8, 9), DNA-PKcs (10), MDM2 (11), CTNNBL1 (12), Spt5 (13), and PTBP2 (14). Unfortunately, however, none of these proteins could show any functional correlation to support the target specificity of AID. There is no clear mechanism to limit the number of target loci. Most of the proteins like RNA polymerase II, protein kinase A, Spt5, and PTBP2 are rather ubiquitous and interact with many proteins other than AID. PTBP2 is a splicing factor, and Spt5 is one of the transcription elongation factors that associate with RNA polymerase II. Replication protein A, DNA-PKcs, and MDM2 are proteins involved in general DNA repair. CTNNBL1 was later shown to be dispensable for CSR (15).

AID has been shown to have the nuclear localization signal and nuclear export signal in its N terminus and C terminus, respectively (16, 17). The deletion of the nuclear localization signal region of AID results in loss of the AID functions for both SHM and CSR (16). A series of mutations at the N teminus of AID also causes defects in CSR as well as SHM (18). Although no mutations at the N terminus of AID have been shown to cause CSR-specific loss of the AID function, some AID mutants with point mutations in the N-terminal region retain substantial CSR activity but severely damage SHM activity, which is most likely due to a combination of partial loss of DNA cleavage activity and less efficient cleavage of the V region compared with S regions (18–20). Conversely, a S3A mutaiton augments both CSR and SHM (21). On the other hand, the deletions and/or mutations in the nuclear export signal region (residues 183–198) result in loss of the AID function for CSR but not SHM, probably because AID with the C-terminal deletion has normal DNA cleavage activity (19, 22). The results suggest that AID has at least two functions: DNA cleavage of V and S region associated with the N-terminal region, and CSR-specific activity associated with the C-terminal region. In addition, the C-terminal region was shown to be responsible for interaction with poly $(A)^+$ RNA (23). We proposed that the C-terminal region of AID might be responsible for generation of recombination synapsis factor (19). Therefore, we assumed that cofactors interacting with the C-terminal region of AID might be responsible for CSR-specific activity rather than DNA cleavage, whereas cofactors interacting with the N-terminal region of AID might be responsible for DNA cleavage of both V and S regions.

In the present study, we screened AID association molecules by yeast two-hybrid screening and coimmunoprecipitation. We then assessed their functional involvement in CSR and SHM.

Author contributions: I.-m.O., K.Y., and T.H. designed research; I.-m.O., K.O., M.K., K.Y., S.K., H.N., Y.K., H.T., T.N., and S.-I.I. performed research; H.N. and R.S. contributed new reagents/analytic tools; I.-m.O. and T.H. analyzed data; and I.-m.O. and T.H. wrote the paper.

The authors declare no conflict of interest.

[1]Present address: Division of Immune Regulation, Institute for Genome Research, University of Tokushima, 770-8503, Japan.

[2]Present address: Drug Discovery Research Laboratories, Kyowa Hakko Kirin Co., Ltd., Shizuoka 411-8731, Japan.

[3]Present address: Department of Medical Oncology, Dana-Farber Cancer Institute, Boston MA 02115.

[4]Present address: Laboratory for Mucosal Immunity, RIKEN Research Center for Allergy and Immunology, Yokohama 230-0045, Japan.

[5]Present address: Department of Bioscience, Nagahama Institute of Bio-Science and Technology, Shiga 526-0829, Japan.

[6]To whom correspondence should be addressed: E-mail: honjo@mfour.med.kyoto-u.ac.jp.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1104423108/-/DCSupplemental.

Because AID associates with a numerous molecules, we used AID mutants as negative controls and chose the association molecules specific to wild-type AID. Among these molecules we identified Spt6, whose interaction was blocked by AID mutations at the N terminus. Knockdown of Spt6 resulted in great reduction of CSR in both the endogenous Ig locus in B cells and an artificial substrate in fibroblasts. Surprisingly, however, knockdown of Spt6 did not reduce but slightly enhanced SHM in an artificial substrate in B cells. These results indicate that Spt6 is involved in differential regulation of CSR and SHM by AID.

Results and Discussion

Proteins Physically Interacting with AID. To identify AID-interacting proteins that may account for the target specificity of AID function in CSR and SHM, we overexpressed mouse (m) and human (h) AID in a mouse B-cell line, CH12F3-2A, which can switch from IgM to IgA upon stimulation. We used AID tagged with GFP-FLAG (GF) at its C terminus to avoid masking the FLAG epitope by putative large AID association proteins. The addition of GF to the C terminus of AID had little effect on the function of AID to induce both CSR and SHM. Cytoplasmic extracts of CH12F3-2A with AID-GF were fractionated by centrifugation through a glycerol density gradient (10–50%, vol/vol). Each fraction collected was analyzed for the presence of AID-GF by GFP fluorescence (Fig. S1*A*). The distribution of AID-GF was broad. Similar broad distribution of endogenous AID was observed in CH12F3-2A extracts using an anti-AID antibody (Fig. S1*B*). However, the increase of the NaCl concentration from 150 mM to 500 mM reduced the overall size distribution and sharpened the distribution profile (Fig. S1*C*). The results indicate that AID interacts with a large number of cytoplasmic proteins, some of which can be removed at 500 mM NaCl. The complex formation was not due to GFP because GF alone formed a small and sharp peak. RNase A treatment reduced the size of the peak only slightly at 150 mM NaCl but hardly at 500 mM NaCl, indicating that there are some AID protein complexes containing RNA but the majority of the AID complex is formed through the protein–protein interaction.

We then immunoprecipitated AID-GF interacting molecules with the anti-FLAG antibody from the cytoplasmic extracts and fractionated by SDS/PAGE, followed by MS. As shown in Fig. 1*A*, a huge variety of proteins were coimmunoprecipitated with AID-GF compared with GF. Therefore, we decided to compare coimmunoprecipitates between AID and its loss-of-function mutant. We used the human AID mutant P13 (M139V) defective in both SHM and CSR activities (22). Similarly diverse proteins with similar intensity were coimmunoprecipitated with the P13 mutant when we used an equal number of cells for wild-type and mutant AID (Fig.1*A*).

Among coimmunoprecipitates, Spt6, Trim28, Nucleolin, Skiv2l2, Zfp84, CRM1, and eEF1α were clearly more abundant in wild-type AID-GF (Fig. 1*A*). We also obtained the following two groups of coimmunoprecipitates from CH12F3-2A cells: (*i*) proteins involved in the nuclear–cytoplasmic transport, including importin 4 and importin β3 (Ranbp 5); and (*ii*) proteins involved in the translation or degradation of proteins, and chaperones, including eEF1α, Hsp70, Stip1, TCP1, CCTq, and KIAA1967. Identification of proteins in the group (*i*) and CRM1 indicates that the coimmunoprecipitations in the current condition is suitable to detect expected functional partners of AID because AID has been reported to be actively exported from nucleus to cytoplasm in a CRM1-dependent manner (16, 17). The degradation-related molecules and chaperons in group (*ii*) were coimmunoprecipitated probably because overexpressed wild-type and mutant AID-GF were targets of the degradation or inactivation machinery. Although eEF1α also showed a striking difference between wild-type AID and P13 mutant, we suspected that eEF1α was directly associated with mRNA to which AID interacts as reported previously (23). In fact the treatment of cell extracts with RNase A before immunoprecipitation significantly reduced eEF1α from the coimmunoprecipitates with AID-GF (Fig. 1*B*).

In a separate series of experiments, we expressed hAID-FLAG (F) and hAID mutants including L172A-F, ΔN10-hAID-F

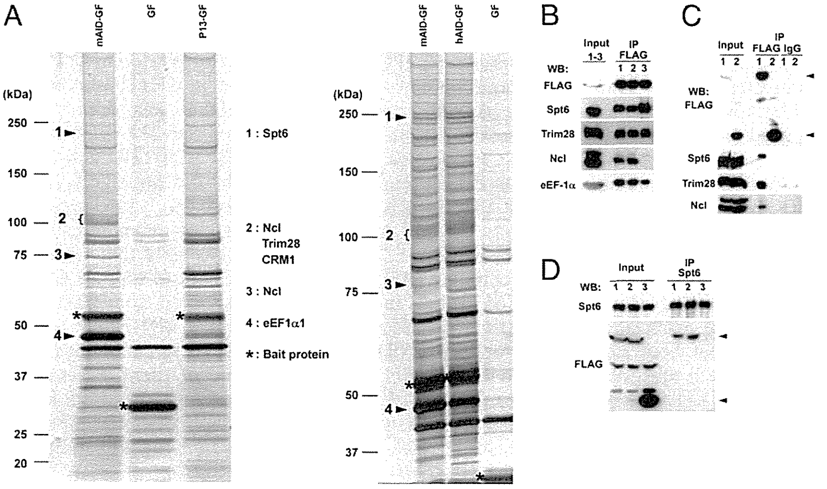


Fig. 1. Identification of AID-interacting proteins. (*A*) Silver staining of proteins coimmunoprecipitated with mouse AID-GFP-FLAG (mAID-GF), human AID-GF (hAID-GF), GF, and P13-GF from cytosolic extracts prepared from CH12F3-2A cells expressing either of the AID-GFs. Bands were excised and analyzed by MS. Proteins that appeared to be obviously more abundant in mAID-GF and hAID-GF are indicated. Nucleolin appears at two positions owing to possible modification. (*B*) Western blot analyses of immunoprecipitates with anti-FLAG M2 antibody from mAID-GF–expressing CH12F3-2A extracts treated with DNase I (lane 2) or RNase A (lane 3). Immunoprecipitates from untreated extracts are shown in lane 1. (*C*) Western blot analyses of immunoprecipitates either with anti-FLAG M2 antibody or with mouse IgG from cytosolic extracts prepared from CH12F3-2A cells expressing mAID-GF (lane 1) or GF (lane 2). Arrows indicate mAID-GF or GF. (*D*) Western blot analyses of immunoprecipitates with anti-Spt6 mAb from cytosolic extracts prepared from CH12F3-2A cells expressing mAID-GF (lane 1), hAID-GF (lane 2), or GF (lane 3). Two nonspecific bands appeared in input by FLAG Western blot.

(N-terminal 10 residue truncation), and JP8Bdel-F (Table S1) in 293T cells and compared proteins immunoprecipitated with the anti-FLAG antibody by MS. We picked up seven proteins that were specifically coimmunoprecipitated with hAID-F in all four repeated experiments. The list of such proteins is shown in Table S2. Surprisingly, none of them overlapped with the above experiments in CH12F3-2A cells. The discrepancy could be at least in part due to the difference in tags to AID and cells used.

Spt6 Is Required for CSR in B Cells and Fibroblasts. We first focused on Spt6, Trim28, and Nucleolin because they are involved in nucleic acid metabolism and most distinctly associated with wild-type AID. We confirmed that Spt6, Trim28, and Nucleolin were coimmunoprecipitated with AID-GF but not GF by Western blotting (Fig. 1*B*). The control IgG did not precipitate any of them (Fig. 1*C*). The association of Spt6 and Trim28 with AID-GF was not reduced by either RNase A or DNase I treatment of the cell extracts before immunoprecipitation, whereas the association of Nucleolin with AID-GF was almost completely abolished by RNase A treatment, suggesting that RNA bridges AID with Nucleolin (Fig. 1*B*). In addition, the association between Spt6 and AID was further confirmed by detection of AID-GF in immunoprecipitates with an anti-Spt6 antibody (Fig. 1*D*).

Interaction of AID with Spt6 was also supported by yeast two-hybrid screening. We screened a human lymph node cDNA library fused to the GAL4 activation domain (AD), with hAID fused to the GAL4 DNA-binding domain (BD) as bait, and a mouse pre B-cell cDNA library fused to the GAL4 AD, with mAID fused to the GAL4 DNA-BD as bait. The C-terminal fragments of hSPT6 (3817–5178 nt) and mSpt6 (4050–5178 nt), which contain Src homology 2 domain, were isolated from the human and mouse libraries, respectively (Fig. S2). This interaction was further confirmed by coimmunoprecipitation of hAID with hSPT6 fragments fused with GST.

We therefore examined the functional involvement of Spt6 in CSR by knocking down its expression in CH12F3-2A cells. Knockdown of Spt6 or AID significantly reduced CSR efficiency (Fig. 2 *A* and *B*). However, 1.5 μg of Spt6 siRNA decreased AID mRNA and germline transcript (GLT) of Cα significantly, although GLT of Cμ was intact (Fig. 2*C*). Therefore, we further examined the effects of Spt6 knockdown using the AID-ER (fusion protein of AID and the hormone-binding domain of the estrogen receptor) system, in which CSR can be induced rapidly upon 4-hydroxy tamoxifen (OHT) addition without de novo transcription and translation of AID. Knockdown of Spt6 with 0.6 μg of siRNA reduced the efficiency of CSR almost in parallel with the degree of Spt6 protein reduction (Fig. 3 *A*–*C*). However, Stp6 knockdown did not affect the amounts of AID-ER protein and GLTs (Fig. 3 *C* and *D*). The results indicate that Spt6 is required for CSR.

To further confirm the involvement of Spt6 in CSR, we used the artificial switch substrate in a mouse fibroblast cell line, NIH 3T3 (24). We knocked down Spt6 using a mixture of siRNAs in NIH 3T3 cells expressing the artificial switch substrate of CSR and AID-ER. siRNAs against Spt6 significantly reduced CSR efficiency in the artificial switch substrate compared with control siRNAs against LacZ (Fig. 4*A*). Consistently, postswitch transcripts were decreased by Spt6 knockdown, although it did not reduce the amounts of AID-ER mRNA and preswitch transcripts Pre-Tr1 and Pre-Tr2 (equivalent to GLTs) (Fig. 4*B*). These results further confirmed that Spt6 is required for CSR.

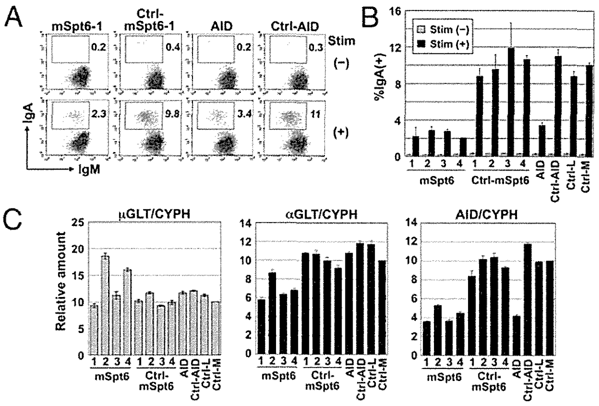


Fig. 2. CSR is inhibited by Spt6 knockdown in CH12F3-2A cells. (*A* and *B*) Spt6 knockdown severely reduced CSR efficiency. CH12F3-2A cells (1.5 × 10^6) were introduced with 1.5 μg of siRNAs against mSpt6, scrambled siRNA for them, an siRNA against mAID, a scrambled siRNA for it, or negative control siRNAs with low (36%) and medium (48%) GC contents (Ctrl-L and Ctrl-M, respectively). The GC contents of oligos mSpt6-1, -2, and -4 are medium (45–55%), and those of oligos mSpt6-3 and AID are low (35–45%). Twenty-four hours after siRNA introduction, cells were stimulated with CD40L, IL-4, and TGF-β for 24 h. The percentages of IgA^+ cells in the live population are indicated. Representative FACS profiles are shown (*A*). The mean ± SD values were obtained from triplicate experiments (*B*). (*C*) siRNAs against Spt6 reduced the amount of αGLT and AID transcripts. Quantitative PCR analyses for μGLT, αGLT, and AID transcripts in Spt6-knockdown cells. Values were normalized by cyclophilin (CYPH). Unstimulated and stimulated cells were analyzed for μGLT and for αGLT and AID transcripts, respectively.

Other Candidates Are Not Required for CSR. We then examined involvement of other AID-binding proteins in CSR by knockdown assay. Knockdown of Trim28 also reduced CSR but simultaneously decreased AID mRNA in CH12F3-2A cells (Fig. S3 *A* and *B*). We thus examined the effect of Trim28 knockdown in the AID-ER system. Trim28 knockdown in the AID-ER system did not affect CSR, suggesting that Trim28 blocked CSR by inhibiting AID transcription (Fig. S3 *C* and *D*). Involvement of Trim28 in the transcriptional regulation of AID was further confirmed by the fact that both AID transcription and CSR are drastically reduced in Trim28-deficient B cells (Fig. S3 *E* and *F*). We concluded that the association of Trim28 with AID is not related to AID function. Knockdown of Trim28 did not show any significant effects on CSR in NIH 3T3 cells either (Fig. 4 *A* and *B*).

Knockdown of Nucleolin did not significantly affect CSR efficiency, either in CH12F3-2A cells or NIH 3T3 cells, although Nucleolin protein was dramatically reduced by knockdown. Knockdown of Skiv2l2 only slightly reduced CSR efficiency, both in CH12F3-2A cells and NIH 3T3 cells, whereas knockdown of Zfp84 did not significantly affect CSR in CH12F3-2A cells. Therefore, we concluded that Nucleolin, Skiv2l2, and Zfp84 do not play major roles in CSR, although we could not exclude the possibility that residual amounts of target proteins were still sufficient to support CSR. We then examined whether candidates identified from 293T cells by specific coimmunoprecipitation with hAID-F are involved in CSR (Table S2). Knockdown of these candidates was carried out in CH12F3-2A cells, but none of them affected CSR significantly except for hnRNPA1, which we could not knockdown. We could therefore identify only Spt6 that has functional relevance for the CSR activity of AID among all candidates detected by physical association with AID.

Spt6 Is Dispensable for SHM in B Cells. We next examined the effect of Spt6 knockdown on SHM in a human B-cell line BL2. To assess SHM efficiency sensitively and quickly, we took advantages of a modified GFP substrate of SHM (Fig. 5*A*). In addition, we used a C-terminal truncation mutant of AID, JP8Bdel, which has stronger SHM activity but marginal CSR activity (19, 22). In this system, OHT-activated JP8Bdel-ER protein caused loss of GFP

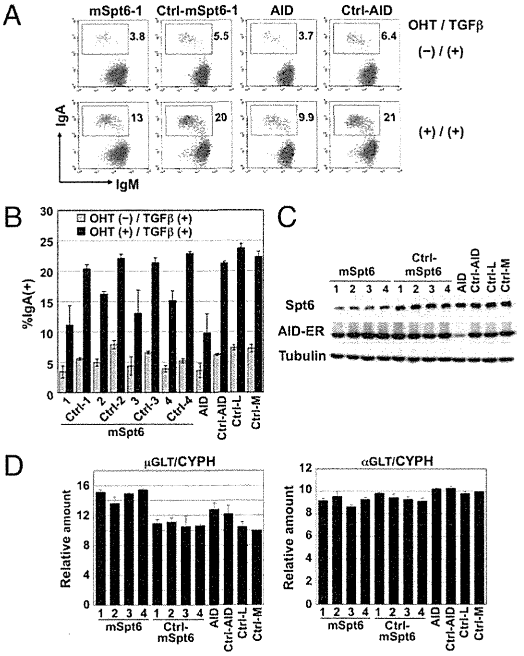


Fig. 3. CSR is inhibited by Spt6 knockdown in mAID-ER–expressing CH12F3-2A cells without affecting the amount of AID. (*A* and *B*) Spt6 knockdown reduced CSR efficiency. CH12F3-2A cells expressing mAID-ER (1.5 x 10^6) were introduced with 0.6 μg of siRNAs against mSpt6, scrambled siRNA for them, an siRNA against mAID, a scrambled siRNA for it, or negative control siRNAs with low and medium GC contents (Ctrl-L and Ctrl-M, respectively). Twenty-four hours after siRNA introduction, cells were stimulated with OHT and TGF-β for 24 h. The percentages of IgA^+ cells in the live population are indicated. Representative FACS profiles are shown (*A*). Mean ± SD values were obtained from triplicate experiments (*B*). (*C*) siRNAs against Spt6 efficiently reduced the amount of Spt6 protein but did not affect the amount of AID-ER protein. (*D*) Quantitative PCR analyses for μGLT and αGLT in Spt6-knockdown cells. Stimulated cells were analyzed. CYPH, cyclophilin.

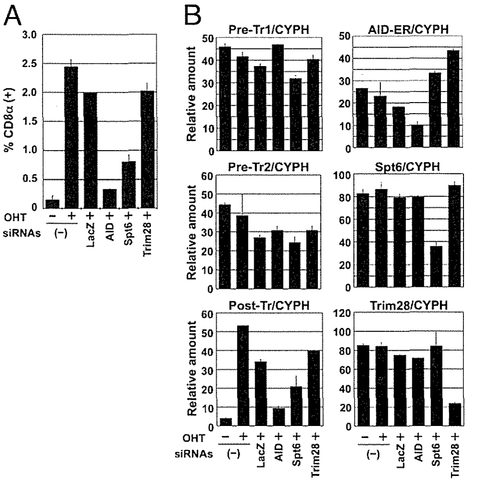


Fig. 4. Spt6 knockdown inhibited CSR in the artificial switch substrate SCI (μ, α) in NIH 3T3 cells. (*A*) NIH 3T3 cells expressing mAID-ER and SCI (μ, α) were introduced with d-siRNAs against mSpt6, mTrim28, mAID, and LacZ together with a DsRed-expressing plasmid as a transfection indicator. Twenty-four hours after transfection, cells were stimulated with OHT for 36 h. The percentages of switched $CD8\alpha^+$ cells in the live and $DsRed^+$ population are indicated. Mean ± SD values were obtained from triplicate experiments. (*B*) Quantitative PCR analyses for pre- (Pre-Tr1 and Pre-Tr2) and postswitch transcripts (Post-Tr), AID-ER, Spt6, and Trim28 transcripts. CYPH, cyclophilin.

fluorescence due to the accumulation of deleterious mutations in the GFP gene. Excessive mutations induced by JP8Bdel caused cell death. AID knockdown by three different siRNA oligos inhibited loss of GFP fluorescence, as well as the accumulation of point mutations in the GFP gene and cell death, confirming that these events were dependent on AID function and thus useful indicators for SHM (Fig. 5 *B* and *C* and Table S3).

Surprisingly, SPT6 knockdown did not inhibit but rather slightly augmented the frequency of GFP-negative cells, as well as actual mutation frequency in the GFP gene and cell death, without affecting the amount of JP8Bdel-ER protein (Fig. 5 *C* and *D* and Table S3). Although the difference was modest, the relative increase of mutation frequencies correlated well with the knockdown efficiency of each oligo against SPT6. It should be noted that both fluorescence and point mutations in the GFP gene were not reduced by SPT6 knockdown in the absence of OHT, indicating that SPT6 knockdown did not affect transcription of the SHM target. These results clearly showed that Spt6 is not required for SHM but rather inhibitory to SHM.

Spt6 Interacts with AID Through Its N Terminus. Because the C-terminal region of Spt6 interacts with AID, we then examined whether a specific region of AID is responsible for the association with Spt6. Deletion of the N-terminal residues 2–26 of AID abolished the association with Spt6, suggesting that this region contains residues involved in the interaction with Spt6 (Fig. 6 *A* and *B* and Table S1). We further tested AID mutants that carry deletion or mutation(s) in the N-terminal region and found that the deletion of residues 2–10 but not residues 2–5 of AID abolished the association with Spt6, suggesting that residues 6–10 may be responsible for the association with Spt6. Amino acid substitution experiments showed that only M6 was critical within residues 6–10 to the association with Spt6. Other mutations in the N-terminal region [G23S, V18S-R19V, W20K, and P7 (R24W)], which have been shown to reduce SHM more drastically than CSR or to abolish both, did not affect the association with Spt6 (Fig. 6*B* and Table S1). In agreement with Fig. 1, P13-GF actually lost the binding capacity with Spt6.

Next, we tried to examine the involvement of the C-terminal region (residues 183–198) of AID in the association with Spt6. To avoid insolubility of the C-terminally deleted (ΔC) AID by the spontaneous accumulation in the nucleus, we combined the deletion or mutation(s) in the N-terminal region with the deletion of the C-terminal region because most N-terminal mutants (P7, V18S-R19V, and W20K) did not accumulate in the nucleus even when nuclear export was blocked by leptomycin B (16, 18). Such N-terminal mutants that additionally lacked the C-terminal region (P7-ΔC, V18S-R19V-ΔC, and W20K-ΔC) still associated with Spt6, indicating that the C-terminal region of AID is not involved in the association with Spt6. A human AID mutant P20 carrying a 34-aa insertion after residue 182 did not show any reduction in the association with Spt6, although P20 has severe defect in CSR but little effect on SHM (22).

Dissociation of AID Mutant Activities from Their Spt6 Binding. Studies on series of AID mutants clearly demonstrated that the C-terminal region of AID is required for CSR-specific function other than DNA cleavage. This function is assumed to be related to synapsis formation of cleaved ends (19). On the other hand, the N-terminal region of AID is required for DNA cleavage of both V and S regions. Because Spt6 is required only for CSR, the interaction of Spt6 with N-terminal residues 6–10 of AID was puzzling. To

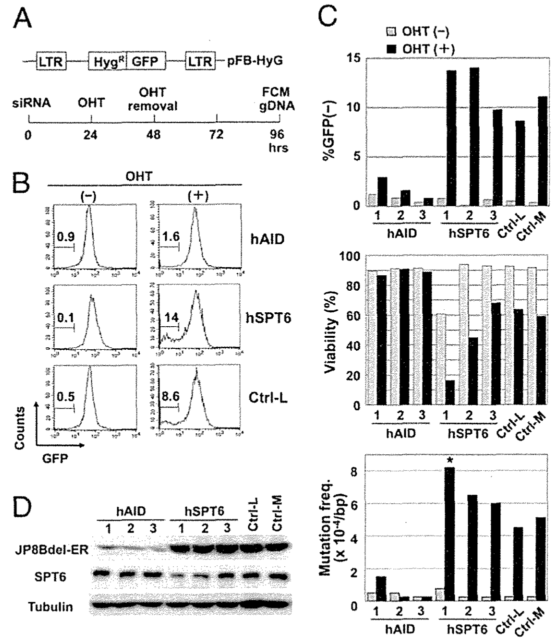


Fig. 5. Spt6 knockdown augmented SHM in BL2 cells. (*A*) Schematic representation of the artificial SHM substrate and the SHM assay procedure. BL2 cells expressing JP8Bdel-ER and the artificial SHM assay substrate were introduced with siRNAs, stimulated for 24 h with OHT, and incubated for an additional 48 h in the absence of OHT. Then cells were harvested for flow cytometry (FCM) and genomic DNA extraction. (*B* and *C*) SPT6 knockdown augmented the SHM efficiency in the artificial substrate. BL2 cells expressing JP8Bdel-ER and the artificial SHM assay substrate (1.5×10^6) were introduced with 3.0 μg of siRNAs against hSPT6, siRNAs against hAID, or negative control siRNAs with low (36%) and medium (48%) GC contents (Ctrl-L and Ctrl-M, respectively). The GC contents of oligos hAID-1, -2, -3, and hSPT6-3 are medium (45–55%) and those of oligos hSPT6-1 and -2 are low (35–45%). The percentages of GFP$^-$ cells are indicated (*B*). Graphical summary of the percentages of GFP$^-$ cells, viability, and mutation frequencies in the GFP sequence (*C*). Statistical significance was evaluated against the corresponding control oligo by χ^2 test. *$P < 0.05$. Data are representative of three independent experiments. (*D*) Two siRNA oligos against SPT6 (1 and 2) reduced the amount of SPT6 protein but did not affect the amount of JP8Bdel-ER protein. Note that the other siRNA oligos against SPT6 (3) did not substantially reduce the amount of SPT6 protein.

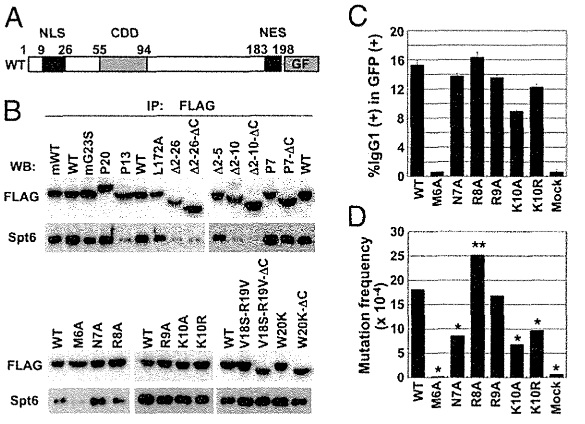


Fig. 6. AID interacts with Spt6 through its N terminus. (*A*) Schematic representation of wild-type and mutant AID-GF constructs. (*B*) Western blot analyses of immunoprecipitates with anti-FLAG from cytosolic extracts of CH12F3-2A cells expressing wild-type AID-GF or mutant AID-GFs. All AID constructs are of human origin except for mWT and mG23S. An equal amount of wild-type and mutant AID-GF proteins was analyzed by adjusting the loading amounts of immunoprecipitates. (*C*) AID-deficient splenocytes were stimulated with LPS for 48 h and infected with retroviruses expressing mutant AID-GFs. Cells were stimulated for additional 48 h in the presence of LPS and IL-4. The percentages of IgG1$^+$ cells in the GFP$^+$ population are indicated. Mean ± SD values were obtained from triplicate experiments. (*D*) NIH 3T3 cells harboring an SHM substrate pI were infected with retroviruses expressing wild-type and mutant AID-GFs and cultured for 7 d. Genomic DNA was extracted for sequencing analysis. GF was used as mock. Statistical significance was evaluated against WT by χ^2 test. *$P < 0.001$, **$P < 0.05$.

monitor the function of the mutants at residues 6–10, each was fused with GFP in the retroviral expression vector and introduced to AID-deficient spleen cells (Fig. 6*C*). We also tested the SHM activities of these mutants in a GFP substrate expressed in NIH 3T3 cells (Fig. 6*D* and Table S4). The point mutation at the residue 6 (M6A) was totally defective for both CSR and SHM. The mutations at residues 7, 9, and 10 reduced SHM as well as CSR, albeit to a less extent. By contrast, the R8A mutant rather augmented CSR and SHM activities.

Although all of R8A, R9A, and K10R mutants had significant modification of their activities, none of them changed the interaction with Spt6 (Fig. 6*B*). Conversely, N7A augmented interaction with Spt6, although N7A reduced both CSR and SHM activities. M6A that abolished Spt6 interaction lost both SHM and CSR, although Spt6 is involved in only CSR. Human AID mutation M6T also lost both CSR and SHM (25). It is possible that M6A mutation altered the gross structure of AID to abolish the DNA cleavage function, resulting in the loss of both CSR and SHM. Although the N-terminal region of AID seems to be responsible for its interaction with the C-terminal region of Spt6, it is not clear whether this interaction is essential for the DNA cleavage function of AID.

How Does Spt6 Differentially Regulate CSR and SHM? The target specificity of known specific recombination is determined by combination of *cis* elements (the DNA sequence/structure) and *trans* elements (DNA binding proteins and the chromatin modification mark of the target locus). In VDJ recombination in the Ig genes, the recombination signal sequence is widely distributed in the genome, but the chromatin modification [i.e., histone3 lysine4 trimethylation (H3K4me3)] recognized by RAG2 is essential to cleave the accurate target (26, 27). In meiotic recombination, Spo11 (topoisomerase II) cleaves at loosely conserved DNA target sequences that are also recognized by zinc finger-histone methyltransferase (PRDM9) to generate H3K4me3 at the target chromatin (28–31). Without PRDM9, meiotic recombination is abortive. We have also shown that H3K4me3 at the target S region is essential for CSR (32). The FACT complex composed of SSRP1 and Sptl6 is a histone chaperone and modulates the histone transmodification cascade. We have shown that the FACT complex is essential for CSR (32). In the absence of FACT, H3K4 trimethyl modifications are reduced at the Sμ and Sα regions, which is associated with S region cleavage defect.

From these studies, it is likely that Spt6 can determine the target specificity of CSR at least by two strategies: (*i*) recognition of DNA sequence or (*ii*) modification of chromatin. Because Spt6 does not bind DNA directly, it is unlikely that Spt6 directly recruits a DNA cleaving enzyme to any DNA region. In addition, Spt6 associates with RNA polymerase II, which binds both V and S regions. Spt6 is thus unlikely to guide AID specifically to S regions. Because Spt6 is another histone chaperone protein, it is important to examine whether Spt6 also affects the histone modification cascade and thus causes defect in CSR. It is also interesting to analyze why Spt6 is slightly inhibitory to SHM. The histone modification cascade in the S region and V region may be different, which triggers interesting possibilities for differential regulation of SHM vs. CSR.

There are several other possible mechanisms whereby Spt6 differentially regulates CSR and SHM. Because Spt6 has also been reported to direct Iws1-dependent mRNA splicing and export (33, 34), CSR-related function of Spt6 could involve mRNA splicing and export. Stp6 is also involved in transcriptional regulation of a large number of genes, some of which may be responsible only for CSR. Further analyses are required to uncover the precise role of Spt6 in regulation of CSR but not SHM.

Materials and Methods

RNA Interference. A diced siRNA (d-siRNA) pool was prepared using BLOCK-iT Complete Dicer RNAi Kit according to the manufacture's instructions (Invitrogen). Primers used to amplify template cDNAs for AID, Trim28, and Spt6 of mouse origin were as follows: mAID-F: 5′-CAA GGG ACG GCA TGA GAC CTA CCT-3′; mAID-R: 5′-TCT CGC AAG TCA TCG ACT TCG TAC-3′; mTrim28-F: 5′-CCA AGG AGG TTC GAA GCT CGA TCC-3′; mTrim28-R: 5′-GGA CCT TCA GTC AGA GGC ATC AAC-3′; mSpt6-F: 5′-CAG CAG TTC CTC TAC GTG CAA ATG-3′; and mSpt6-R: 5′-ACT GGA TCA AGG CCT GGC TGT AAG-3′. Stealth siRNAs were introduced into CH12F3-2A or BL2 cells using Amaxa Nucleofector (Amaxa Biosystems). Stealth siRNAs were purchased from Invitrogen: mSpt6-1, -2, -3, and -4 (MSS209819, 209820, 209821, NM_009297_stealth_3806), AID (MSS235859), hSPT6-1, -2, and -3 (HSS110374, 110375, 110376), hAID-1, -2, and -3 (HSS126211, 126212, 126213), mTrim28-1, -2, -3, and -4 (MSS211796, 211797 211798, NM_011588_stealth_1165), and hTRIM28-1, -2, -3, and -4 (HSS115468, 115470, NM_005762_stealth_883, NM_005762_stealth_2386). The efficiency of nucleofection was confirmed to be more than 90% by introducing fluorescein-labeled siRAN oligo.

CSR Assay. CH12F3-2A cells were stimulated for 24 h with CD40L, TGF-β, and IL-4 24 h after introducing siRNA. The surface expression of IgM and IgA was analyzed by staining cells with FITC-conjugated anti-mouse IgM (Southern Biotechnology Associates) and PE-conjugated anti-mouse IgA (Southern Biotechnology Associates). Flow cytometric analyses were performed with a FACSCalibur, and data were analyzed by CellQuest software (BD Biosciences). Live cells were selected for the analyses by forward- and side-scatter intensity and propidium iodide (PI) gatings. CH12F3-2A cells expressing AID-ER were stimulated with OHT and TGF-β for 24 h after introducing siRNA, and the surface expression of IgM and IgA was analyzed by flow cytometry as described above. NIH 3T3 cells expressing the artificial switch substrate SCI(μ, α) and mAID-ER were introduced with d-siRNA and a DsRed-expressing plasmid as a transfection indicator using Lipofectamine 2000 (Invitrogen). Twenty-four hours after transfection, cells were stimulated with OHT 24 for 36 h and stained with allophyco-cyanin–conjugated anti-mouse CD8α (eBioscience). The amounts of μGLT and αGLT were evaluated by quantitative PCR as described previously (32).

SHM Assay. The hygromycin phosphotransferase and EGFP cDNA were fused in-frame in pFB to generate an artificial SHM substrate, pFB-HyGFP. BL2 cells were introduced with AID JP8Bdel-ER and pFB-HyGFP by retroviral infection. A clone expressing AID JP8Bdel and pFB-HyGFP was chosen after selection with puromycin and hygromycin. The clone was stimulated for 24 h with OHT 24 h after introducing siRNA and incubated for an additional 48 h in the absence of OHT. Expression of GFP and survival were evaluated by flow cytometry. Live cells were selected for the analyses by forward- and side-scatter intensity and PI gatings. Genomic DNA was extracted, and GFP sequence was amplified and analyzed. NIH 3T3 cells harboring a SHM substrate pI were infected with retroviruses expressing wild-type and mutant AID-GFs and cultured for 7 d. Genomic DNA was extracted and GFP sequence was amplified and analyzed.

ACKNOWLEDGMENTS. We thank Dr. H. Handa for anti-Spt6 mAb, Dr. Watanabe for pACT2 mouse pre-B-cell cDNA library, Dr. P. Chambon for Trim28$^{f/f}$ mice, and Ms. T. Toyoshima for her excellent technical assistance. This work was supported by Grant-in-Aid 17002015 for Specially Promoted Research from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

1. Kinoshita K, Honjo T (2001) Linking class-switch recombination with somatic hypermutation. *Nat Rev Mol Cell Biol* 2:493–503.
2. Honjo T, Kinoshita K, Muramatsu M (2002) Molecular mechanism of class switch recombination: Linkage with somatic hypermutation. *Annu Rev Immunol* 20:165–196.
3. Muramatsu M, Nagaoka H, Shinkura R, Begum NA, Honjo T (2007) Discovery of activation-induced cytidine deaminase, the engraver of antibody memory. *Adv Immunol* 94:1–36.
4. Pasqualucci L, et al. (2001) Hypermutation of multiple proto-oncogenes in B-cell diffuse large-cell lymphomas. *Nature* 412:341–346.
5. Liu M, et al. (2008) Two levels of protection for the B cell genome during somatic hypermutation. *Nature* 451:841–845.
6. Nambu Y, et al. (2003) Transcription-coupled events associating with immunoglobulin switch region chromatin. *Science* 302:2137–2140.
7. Chaudhuri J, Khuong C, Alt FW (2004) Replication protein A interacts with AID to promote deamination of somatic hypermutation targets. *Nature* 430:992–998.
8. Basu U, et al. (2005) The AID antibody diversification enzyme is regulated by protein kinase A phosphorylation. *Nature* 438:508–511.
9. Pasqualucci L, Kitaura Y, Gu H, Dalla-Favera R (2006) PKA-mediated phosphorylation regulates the function of activation-induced deaminase (AID) in B cells. *Proc Natl Acad Sci USA* 103:395–400.
10. Wu X, Geraldes P, Platt JL, Cascalho M (2005) The double-edged sword of activation-induced cytidine deaminase. *J Immunol* 174:934–941.
11. MacDuff DA, Neuberger MS, Harris RS (2006) MDM2 can interact with the C-terminus of AID but it is inessential for antibody diversification in DT40 B cells. *Mol Immunol* 43:1099–1108.
12. Conticello SG, et al. (2008) Interaction between antibody-diversification enzyme AID and spliceosome-associated factor CTNNBL1. *Mol Cell* 31:474–484.
13. Pavri R, et al. (2010) Activation-induced cytidine deaminase targets DNA at sites of RNA polymerase II stalling by interaction with Spt5. *Cell* 143:122–133.
14. Nowak U, Matthews AJ, Zheng S, Chaudhuri J (2011) The splicing regulator PTBP2 interacts with the cytidine deaminase AID and promotes binding of AID to switch-region DNA. *Nat Immunol* 12:160–166.
15. Han L, Masani S, Yu K (2010) Cutting edge: CTNNBL1 is dispensable for Ig class switch recombination. *J Immunol* 185:1379–1381.
16. Ito S, et al. (2004) Activation-induced cytidine deaminase shuttles between nucleus and cytoplasm like apolipoprotein B mRNA editing catalytic polypeptide 1. *Proc Natl Acad Sci USA* 101:1975–1980.
17. McBride KM, Barreto V, Ramiro AR, Stavropoulos P, Nussenzweig MC (2004) Somatic hypermutation is limited by CRM1-dependent nuclear export of activation-induced deaminase. *J Exp Med* 199:1235–1244.
18. Shinkura R, et al. (2004) Separate domains of AID are required for somatic hypermutation and class-switch recombination. *Nat Immunol* 5:707–712.
19. Doi T, et al. (2009) The C-terminal region of activation-induced cytidine deaminase is responsible for a recombination function other than DNA cleavage in class switch recombination. *Proc Natl Acad Sci USA* 106:2758–2763.
20. Wei M, et al. (2011) Mice carrying a knock-in mutation of Aicda resulting in a defect in somatic hypermutation have impaired gut homeostasis and compromised mucosal defense. *Nat Immunol* 12:264–270.
21. Gazumyan A, et al. (2011) Amino-terminal phosphorylation of activation-induced cytidine deaminase suppresses c-myc/IgH translocation. *Mol Cell Biol* 31:442–449.
22. Ta VT, et al. (2003) AID mutant analyses indicate requirement for class-switch-specific cofactors. *Nat Immunol* 4:843–848.
23. Nonaka T, et al. (2009) Carboxy-terminal domain of AID required for its mRNA complex formation in vivo. *Proc Natl Acad Sci USA* 106:2747–2751.
24. Okazaki IM, Kinoshita K, Muramatsu M, Yoshikawa K, Honjo T (2002) The AID enzyme induces class switch recombination in fibroblasts. *Nature* 416:340–345.
25. Durandy A, Peron S, Taubenheim N, Fischer A (2006) Activation-induced cytidine deaminase: structure-function relationship as based on the study of mutants. *Hum Mutat* 27:1185–1191.
26. Matthews AG, et al. (2007) RAG2 PHD finger couples histone H3 lysine 4 trimethylation with V(D)J recombination. *Nature* 450:1106–1110.
27. Liu Y, Subrahmanyam R, Chakraborty T, Sen R, Desiderio S (2007) A plant homeodomain in RAG-2 that binds Hypermethylated lysine 4 of histone H3 is necessary for efficient antigen-receptor-gene rearrangement. *Immunity* 27:561–571.
28. Wahls WP, Davidson MK (2010) Discrete DNA sites regulate global distribution of meiotic recombination. *Trends Genet* 26:202–208.
29. Parvanov ED, Petkov PM, Paigen K (2010) Prdm9 controls activation of mammalian recombination hotspots. *Science* 327:835.
30. Baudat F, et al. (2010) PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science* 327:836–840.
31. Myers S, et al. (2010) Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. *Science* 327:876–879.
32. Stanlie A, Aida M, Muramatsu M, Honjo T, Begum NA (2010) Histone3 lysine4 trimethylation regulated by the facilitates chromatin transcription complex is critical for DNA cleavage in class switch recombination. *Proc Natl Acad Sci USA* 107: 22190–22195.
33. Yoh SM, Cho H, Pickle L, Evans RM, Jones KA (2007) The Spt6 SH2 domain binds Ser2-P RNAPII to direct Iws1-dependent mRNA splicing and export. *Genes Dev* 21:160–174.
34. Yoh SM, Lucas JS, Jones KA (2008) The Iws1:Spt6:CTD complex controls cotranscriptional mRNA biosynthesis and HYPB/Setd2-mediated histone H3K36 methylation. *Genes Dev* 22:3422–3434.

J. Biochem. 2011;150(1):73–81 doi:10.1093/jb/mvr042

JB THE JOURNAL OF BIOCHEMISTRY

Statistical analysis of features associated with protein expression/solubility in an *in vivo Escherichia coli* expression system and a wheat germ cell-free expression system

Received October 31, 2010; accepted March 3, 2011; published online April 9, 2011

Shuichi Hirose[1,*], Yoshifumi Kawamura[2], Kiyonobu Yokota[1], Toshihiro Kuroita[3], Tohru Natsume[4], Kazuo Komiya[5], Takeshi Tsutsumi[5], Yorimasa Suwa[5], Takao Isogai[5], Naoki Goshima[4] and Tamotsu Noguchi[1]

[1]Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), Tokyo 135-0064; [2]Japan Biological Informatics Consortium (JBiC), Tokyo 135-8073; [3]Toyobo Co., Ltd., Tsuruga Institute of Biotechnology, Fukui 914-0074; [4]Biomedicinal Information Research Center (BIRC), National Institute of Advanced Industrial Science and Technology (AIST), Tokyo 135-0064; and [5]Reverse Proteomics Research Institute, Co., Ltd., Tokyo 110-0044, Japan

*Shuichi Hirose, AIST Tokyo Waterfront Bio-IT Research Building, 2-4-7, Aomi, Koto-ku, Tokyo 135-0064, Japan. Tel: +81-3-3599-8730, Fax: +81-3-599-8081, email: hirose-shuichi@aist.go.jp

Recombinant protein technology is an important tool in many industrial and pharmacological applications. Although the success rate of obtaining soluble proteins is relatively low, knowledge of protein expression/solubility under 'standard' conditions may increase the efficiency and reduce the cost of proteomics studies. In this study, we conducted a genome-scale experiment to assess the overexpression and the solubility of human full-length cDNA in an *in vivo Escherichia coli* expression system and a wheat germ cell-free expression system. We evaluated the influences of sequence and structural features on protein expression/solubility in each system and estimated a minimal set of features associated with them. A comparison of the feature sets related to protein expression/solubility in the *in vivo Escherichia coli* expression system revealed that the structural information was strongly associated with protein expression, rather than protein solubility. Moreover, a significant difference was found in the number of features associated with protein solubility in the two expression systems.

Keywords: *Escherichia coli*/protein expression/protein solubility/statistical analysis/wheat germ cell-free.

Abbreviations: cDNA, complementary DNA; ORF, open reading frame; RF, random forest; SDS–PAGE, sodium dodecyl sulfate–poly-acrylamide gel electrophoresis; SVM, support vector machine.

Obtaining highly concentrated, soluble proteins' preparations is necessary for conducting various structural and biophysical studies or using proteins as materials for pharmaceutical or industrial products. *Escherichia coli*, which is easy to handle and manipulate genetically, is the preferred host for overexpressing recombinant proteins, since it can be cultivated rapidly and inexpensively. Moreover, it generally yields high levels of recombinant proteins (*1*). Since the proteins are expressed by the host, one reason for non-expression is a deleterious interaction with the host's metabolism. In addition, a common reason for insolubility is the formation of inclusion bodies. Therefore, the success rate for obtaining soluble proteins is relatively low. For that reason, the construction of protein overexpression systems is an important experimental challenge.

To overcome these unfavourable circumstances, several solutions have been proposed, based on the results of experimental studies: using a different strain of *E. coli*; modifying the N-terminal (*2*) and C-terminal sequences (*3*); fusion with solubility enhancing tags (*4*) and coexpression with molecular chaperones (*5*). Similarly, various alternative cell-based expression systems have been developed. Such systems utilize yeast, insect cells or murine myeloma cells as hosts (*6*). In recent years, cell-free methods for protein synthesis with extracts from prokaryotic (*7*) or eukaryotic (*8*) cells have become an alternative to cell-based methods. The distinctive feature is an *in vitro* translation system. Cell-free expression systems are popular in proteomics and biotechnology, because of their high levels of protein expression and ease of handling (*9*, *10*).

In theoretical computational science, clear sequence differences between proteins that remain soluble and those that form inclusion bodies have been reported, thereby yielding some successes in predicting protein solubility solely from amino acid sequences (*11*–*17*). The first attempt to determine the interconnection between amino acid sequences and protein solubilities was performed by Wilkinson and Harrison (*11*). They observed that protein solubility is strongly associated with the charge average and the turn-forming residue fraction. Subsequent studies revealed several factors associated with protein expression and solubility. Such knowledge under 'standard' conditions may provide a clue for determining priority targets in a large-scale proteomics analysis. However, the difference between the factors related to protein expression and solubility has remained unclear for genome-scale analyses. In addition, no study has revealed the differences in these factors among protein expression systems.

In this study, we estimated and compared the minimal sets of sequence and structural features associated with protein expression/solubility in an *in vivo E. coli* expression system and a wheat germ cell-free expression system, from an analysis of genome-wide experimental data. The results provide useful information for proteomics analyses.

Materials and Methods

Protein expression experiments

We conducted genome-scale experiments that assessed the overexpression and the solubility of human full-length complementary DNA (cDNA) in an *in vivo E. coli* expression system and a wheat germ cell-free expression system.

In the *E. coli* expression system, the 17,739 human open reading frames [ORFs as Gateway entry clones (*9*)] were first subcloned into the pDEST17 vector (T7 promoter, amino-terminal His-tag fusion), using LR Clonase (Invitrogen). *E. coli* BL21 star (DE3) pLysS was transformed with the reaction products. The SOC expression mixture was then plated on LB agar plates containing ampicillin. For each ORF, LB medium was inoculated with a single colony and grown overnight at 37°C. The overnight culture was diluted 1:100 into SB medium, grown at 37°C for 3 h and cooled to 20°C. Protein expression was then induced by adding isopropyl 1-thio-β-D-galactopyranoside to a final concentration of 0.1 mM. After 16 h at 20°C, the cells were harvested and suspended in BugBuster (Novagen Inc.). Some of the lysate was reserved as the whole cell sample. The lysate was centrifuged at 15,000*g* for 5 min. The supernatants were prepared as the soluble fraction samples. These protein samples were denatured with SDS-sample buffer and fractionated by 12.5% sodium dodecyl sulfate–poly-acrylamide gel electrophoresis (SDS–PAGE). Proteins were visualized with CBB R-250. Protein expression and solubility were judged by visual inspection, according to whether a clear band was detectable at the expected position calculated from the molecular weight in the whole cell and soluble fraction samples. When a clear band in the whole cell sample lane was obviously observed in the presence of bands derived from *E. coli*, it was regarded as overexpression. Conversely, when a specific band was not detected at the expected position, it was regarded as no expression. In the case of a smeared band, it was regarded as low expression. On the same basis, the protein solubility was evaluated using the band in the soluble fraction sample lane.

In the wheat germ cell-free expression system, we expressed 8850 human proteins, for which the entry clones were chosen randomly from our human Gateway entry clone resources (*9*, *18*). The expressed proteins were fused with a carboxy-terminal His-tag (destination vectors: pEW-3H) and were labelled with radioactive amino acids (^{14}C-Leu or ^{35}S-Met). The wheat germ extract was purchased from Toyobo and Cell Free Sciences Co., Ltd. The expressed proteins were separated into soluble and insoluble fractions by centrifugation at 19,000*g* for 20 min. The samples were separated by SDS–PAGE, and the protein expression and solubility were measured by detecting the specific activities of the ^{14}C-Leu and ^{35}S-Met radioisotopes (RI), using a BAS 2000 scanner (Fuji). The rate of dissolution was estimated as described below

$$\text{Rate of dissolution (\%)} = (\text{signal intensity of soluble fraction}/\text{signal intensity of whole sample}) \times 100 \quad (1)$$

When the rate of dissolution was >60%, the protein was regarded as being highly soluble. In contrast, a protein with the rate of dissolution of <40% was regarded as an insoluble protein.

Data sets

We prepared two data sets, based on the number of experiments per sequence, for the statistical analysis. One group, designated as 'data set_S' ('S' means single), comprised sequences for which the protein expression and solubility were experimentally assessed one time. The other group, designated as 'data set_M' ('M' means multiple), comprised sequences for which the experiments were conducted two or more times.

The two data sets were constructed as follows (Fig. 1). First, the clones bearing the same sequence among the experimental expression data were selected (Fig. 1, Step1). The redundancy checks were executed on the nucleotide level and the amino acid level for protein expression and solubility, respectively. If the same sequences existed, they were regarded as one sequence. In the case of protein expression in the *in vivo E. coli*, 474 out of 17,739 clones were redundant, and 227 sequences remained after checking the sequence redundancy. Next, the sequences for which the experimental results were not the same were removed, to use only the reproducible data (Fig. 1, Step2). In the case of protein expression in the *in vivo E. coli*, 44 sequences were removed. The representative sequences were

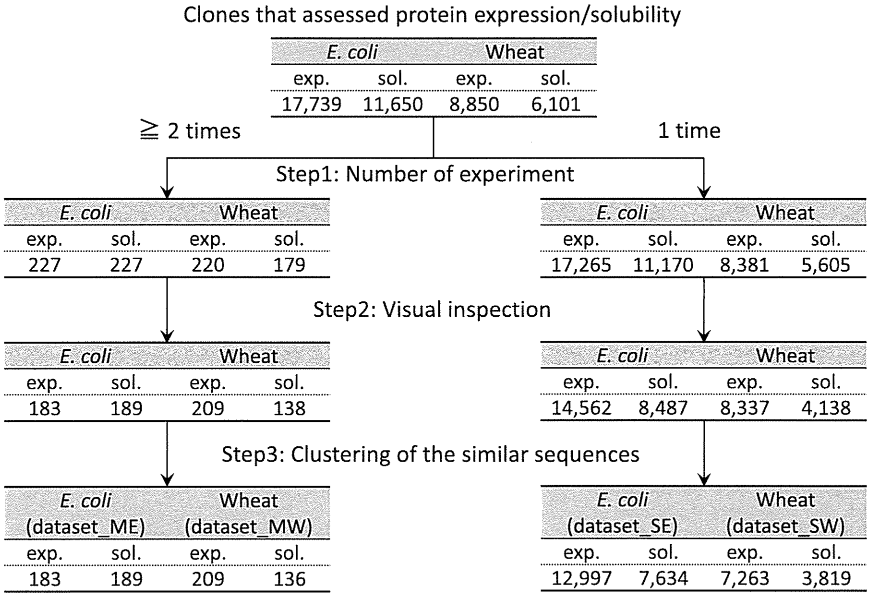


Fig. 1 Construction process of data set. Each table represents the number of data set size in each step.

Table I. Data set sizes for statistical analyses.

		Expression		Solubility	
Expression system	Data set	Positive (%)	Negative (%)	Positive (%)	Negative (%)
E. coli	Data set_ME	113 (61.7)	70 (38.3)	71 (37.6)	118 (62.4)
	Data set_SE	7631 (58.7)	5366 (41.3)	2725 (35.7)	4909 (64.3)
Wheat Germ	Data set_MW	208 (99.5)	1 (0.5)	86 (63.2)	50 (36.8)
	Data set_SW	7062 (97.2)	201 (2.8)	2653 (69.5)	1166 (30.5)

Numbers in parentheses signify ratios of positive data and negative data for respective data sets.

then selected in each data set, to avoid the bias of similar sequences (Fig. 1, Step3). The sequences with pair-wise sequence identity of >80%, using CD-hit (*19*), having similar length >80% were assumed to be in a cluster. The longest sequence in each cluster was selected as the representative sequence of each cluster. This collection of sequences was defined as data set_M. On the other hand, data set_S was constructed from the data from which the redundant clones from the expression data had been removed. In the case of protein expression in the *in vivo E. coli*, 17,265 (=17,739–474) sequences were used. The data that showed a smeared band were removed by visual inspection (Fig. 1, Step2) (see 'Results' section). In the case of protein expression in the *in vivo E. coli*, 2703 sequences were removed. Next, in the same manner as for data set_M, the representative sequences were selected from each cluster consisting of similar sequences (Fig. 1, Step3). This collection of sequences was defined as data set_S. The data set size is shown in Table I.

In this study, data set_M was used for estimating the features associated with the protein expression and solubility; data set_S was used for assessing whether a set of selected features corresponds to the general characteristics on a genomic scale. The initial letter of the expression system was added to the end of the data set name, to distinguish them. For example, 'data set_SE' consists of the sequences for which experimental evaluations were performed one time in the *in vivo E. coli* expression system.

Estimation of the features associated with protein expression/solubility

We defined 437 features to investigate the factors associated with protein expression/solubility in the two kinds of expression systems. The features were divided into two groups, based on the information used for producing them, except for the size of the polypeptide.

The first group was derived from sequence information, from both the nucleotides and amino acids. The nucleotide information included the occurrence frequencies of four single nucleotides, 64 codons and the GC contents. Similarly, the amino acid information contained the occurrence frequencies of 20 single amino acids and the property groups, defined by their chemical properties (eight groups: [GALVI][FYW][ST][DE][NQ][RHK][CM][P]) and physical properties (five groups: [GAVLIP][FWY][STCMNQ][DE][RKH]) (Supplementary Table SI). Additionally, the repeat was defined as the maximum number of consecutive same amino acids or property groups. The values of these features were computed for the entire chains and both terminal regions, defined as 60 bases (meaning 20 amino acid residues), because modification of the terminal regions influences protein expression and solubility (*2–4*). The use of a His-tag fusion raises the possibility that the features in the N-terminal region of the *in vivo E. coli* expression system and the C-terminal region in the wheat germ cell-free expression system may not be evaluated properly. We considered the His-tag to have the same influence on any sequences, since we conducted the protein expression experiments under the same conditions. Therefore, we evaluated them under this hypothesis. In total, the first group was composed of 396 features.

The second group was derived from structural information, obtained with several prediction using amino acid information. The structural information included the secondary structures—α-helix, β-sheet and others predicted by PHD (*20*)—along with the transmembrane regions [predicted using TMHMM (*21*)] and the disordered regions [predicted using POODLE-L (*22*)]. For the secondary structures, the ratio of each element to the entire chain was computed. For the disordered regions, their number of occurrences, lengths and proportions in relation to the entire chain were computed. For the transmembrane regions, the number of occurrences in the entire chain was computed. The structure information also included the occurrence frequencies of single amino acids and the same property groups on the protein surface. The accessible surface area was predicted using RVPnet (*23*). In total, the second group included 40 features.

We estimated which features are associated with protein expression/solubility by analyzing data set_ME and data set_MW. For all features, the statistical difference between positive and negative data was determined using the Student's *t*-test. The positive data of protein expression and solubility mean that a clear band was found in the whole cell sample and the soluble fraction sample. The negative data signify the opposite. A difference of $P<0.05$ was considered significant.

Assessing the generality of the features

To evaluate whether the set of features selected in the previous section corresponds to the general characteristics of protein expression/solubility on a genome-scale in the two expression systems, we built a statistical model that distinguishes between overexpression and low expression, using sequence information only. Similarly, a statistical model to discriminate between soluble and insoluble proteins was built as well. In this study, we applied the random forest (RF) algorithm (*24*) to produce the statistical models.

First, the sequence in the training and evaluation data set was expressed as a multi-dimensional vector that defined the selected features in the previous section as descriptors. The numbers of elements in a vector were 64 and 45 for protein expression and solubility in the *in vivo E. coli* expression system, respectively (see 'Results' section). In contrast, the sequence was expressed by 32 elements in the wheat germ cell-free expression system (see 'Results' section). The statistical models were then built by training data sets. The default values were used as the RF parameters.

The classification abilities of the statistical models for both expression systems were estimated, using two kinds of evaluation methods. One method was a 5-fold cross validation test using data set_M only. The other method was an expanded test. The statistical models were constructed using data set_M. The classification abilities of these models were then estimated, using data set_S. Finally, the classification abilities obtained from the two evaluation methods were compared. Moreover, in order to validate the features, the classification abilities of these models were compared with that of the Wilkinson and Harrison model (*11*). The model was used to predict the *in vivo* solubility of recombinant proteins in *E. coli*:

$$\mathrm{CV} = 15.43\frac{\mathrm{N}+\mathrm{G}+\mathrm{P}+\mathrm{S}}{n} - 29.56\left|\frac{(\mathrm{R}+\mathrm{K})-(\mathrm{D}+\mathrm{E})}{n} - 0.03\right| + 1.71,$$

where N, G, P, S, R, K, D and E are the numbers of asparagines, glycines, prolines, serines, arginines, lysines, aspartic acids and glutamic acids, respectively and n is the total number of residues in the sequence. If $\mathrm{CV}<0$, then the protein is predicted to be soluble. If $\mathrm{CV}>0$, then the protein is predicted to be insoluble.

Results

Comprehensive assessment of protein expression/solubility of human full-length cDNA in two expression systems

The human full-length cDNA was expressed in the two expression systems. The results were analysed

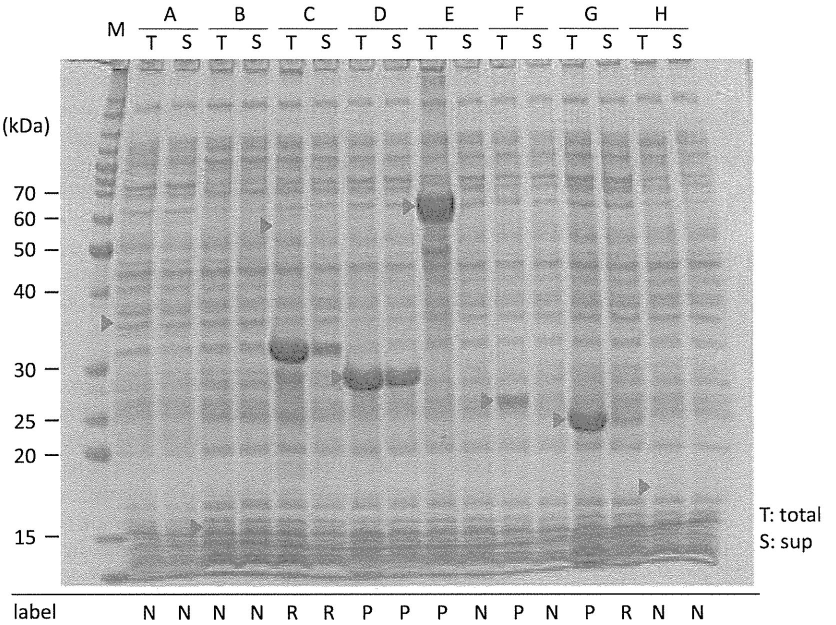

M
A
B
C
D
E
F
G
H
T S T S T S T S T S T S T S T S
(kDa)
70
60
50
40
30
25
20
15
T: total
S: sup
label
N N N N R R P P P N P N P R N N

Fig. 2 Example of an SDS–PAGE analysis for eight proteins expressed in the *in vivo E. coli* expression system. M and A∼H, respectively show molecular weight markers and samples. The T and S lanes, respectively show samples obtained from whole cell samples and soluble fraction samples. The red triangles represent the expected positions calculated from the molecular weights. P and N in the label signify positive and negative data, respectively. R in the label shows data removed from statistical analyses.

using SDS–PAGE (Fig. 2). The gels containing the fractionated proteins expressed in the wheat germ cell-free expression system can be seen at the site HGPD (http://riodb.ibase.aist.go.jp/hgpd/cgi-bin/index.cgi) (*18*). When a clear band is present at the expected position calculated from the molecular weight, such as in lane T of sample D in Fig. 2, the data are considered to be positive. However, when an expected band in SDS–PAGE cannot be detected, such as that in lane T of sample A in Fig. 2, the data are considered to be negative. Data were removed from the following analysis if a smeared band (lane S of sample G in Fig. 2) was observed or a clear band existed at an unexpected position (lane T of sample C in Fig. 2), in order to avoid ambiguity in the experimental data. When the SDS–PAGE results were visually inspected to check the protein expression in the *in vivo E. coli* expression system, 44 of 227 raw data sets were removed between the multiple measurements. Similarly, 16.7 and 23.0% of the raw data were excluded, respectively, from the protein solubility in the *in vivo E. coli* and the wheat germ cell-free expression systems.

The sizes of data set_ME and data set_MW are smaller than those of data set_SE and data set_SW (Table I), but data set_ME and data set_MW are more reliable experimental data. In the *in vivo E. coli* expression system, ~60 and 35% of the proteins, respectively, were expressed and soluble. In contrast, almost all of the proteins were expressed in the wheat germ cell-free expression system: ~65% of the proteins were soluble (Table I). The wheat germ cell-free expression system exhibited higher performance in obtaining soluble proteins. For the wheat germ cell-free expression system, only the protein solubility data were used in the following statistical analyses.

Estimation of the features associated with protein expression/solubility in the two expression systems

The sizes of the polypeptides used to assess the protein expression/solubility experimentally in the *in vivo E. coli* expression system were investigated (Fig. 3A). The average size of the overexpressed polypeptides was significantly longer ($P < 0.05$) than that of the polypeptides with low expression, but no statistically significant difference was found between the sizes of the soluble and insoluble polypeptides. Conversely, in the wheat germ cell-free expression system, the average size of the insoluble polypeptides was significantly longer than that of the soluble polypeptides (data not shown).

Similarly, some sequence and structural features associated with protein expression/solubility were identified from statistical analyses of data set_ME and data set_MW (Fig. 4). In this study, data set_M is not suitable for analyzing the nucleotide information associated with the protein solubility, because data set_M of solubility included sequences that are not identical on the nucleotide level. Consequently, the analysis of the nucleotide information was performed only for the protein expression.

From the perspective of nucleotide information, no GC content or single nucleotide was selected in the *in vivo E. coli* expression system, but 18 out of 61 codons were chosen to have significant contribution to protein expression. Only three rare-frequency codons in the *E. coli* genetic code, among eight tested, passed the Student's *t*-test having significant

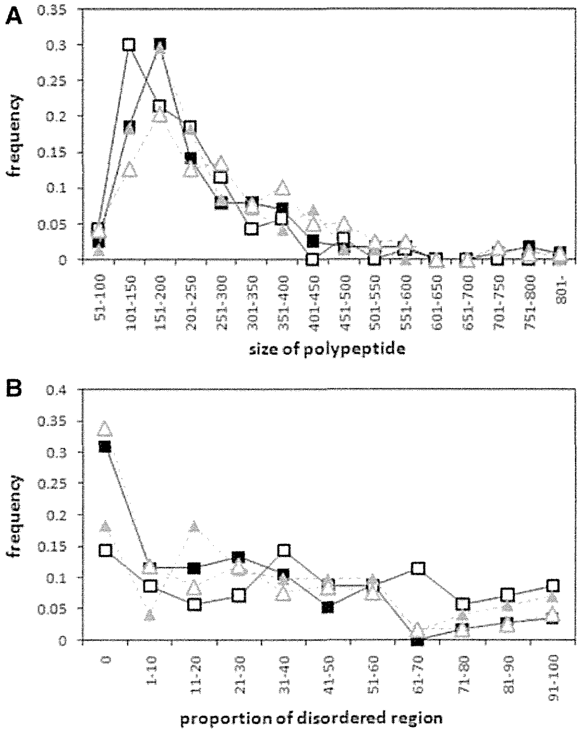


Fig. 3 Distribution of (A) polypeptide sizes and (B) disordered regions in data set_ME. The black squares and grey triangles signify protein expression and solubility, respectively. The filled symbols represent positive data, whereas the open symbols show negative data. The horizontal and vertical axis, respectively show the size and the frequency of the polypeptides.

effect on protein expression. Although it has been suggested that the codon usage influences protein expression (*25–27*), little correlation between rare codons and protein expression was detected in this study. The discrepancy might be explained by the fact that the data set does not include point mutation experiments that change low-usage codons into high-usage ones. Therefore, the estimation suggests the possibility that the influence of rare codons cannot be evaluated. In addition, many selected features are corresponding to amino acids that are encoded by several codons. These observations are the same as those reported by Welch *et al.* (*28*).

Regarding the amino acid information, in the *in vivo E. coli* expression system, the number of features that passed the Student's *t*-test is larger for protein solubility than for protein expression. Particularly, there were many features related to protein solubility in the C-terminal region. Charged residues have a positive effect on both protein expression and solubility, but aromatic residues have a negative effect. In addition, a sulfur-containing residue influences only the protein solubility. In the wheat germ cell-free expression system, the number of selected features is smaller than that in the *in vivo E. coli* expression system. Specifically, the presence of the charged and sulfur-containing residues has little effect on protein solubility. Non-polar residues show the opposite effect in the *in vivo E. coli* expression system.

Regarding the structural information, in the *in vivo E. coli* expression system, the number of features that passed the Student's *t*-test is larger for protein expression than for protein solubility. Statistical analyses revealed that the difficulty of expressing a protein tends to increase in the presence of more disordered regions (Fig. 3B). In contrast, the secondary structure has no effect on protein expression/solubility.

In the wheat germ cell-free expression system, the number of structural features that passed the Student's *t*-test is smaller than that in the *in vivo E. coli* expression system, along with the number of sequence features. In this study, we also examined the correlation between the protein expression/solubility and the number of folded domains predicted by DOMpro (*29*). No significant relation was found (data not shown). This is because more than half of the proteins in our data set have multiple domains, and it is difficult to estimate the number of domains from amino acid information. For that reason, our data set might be unsuitable for analyzing the relation between the number of domains and the protein expression/solubility.

In this study, the definition of the terminal region was 60 nt. To lend credence to the analysis, we estimated the important parameters using new definitions of the terminal region, 30 and 90 nt, and compared them. A strong relationship between protein expression and the presence of rare-frequency codons was not detected in the *in vivo E. coli* expression system, although the some of codons having statistically significant difference changed depending on the length of the terminal region. For the amino acid information, similar features passed the Student's *t*-test. Overall, the results indicated that the tendencies of the important features were the same under any conditions (Supplementary Fig. S1).

Generality of the features

To assess the generality of the features selected in the previous section, we built statistical models that classified the overexpressed proteins and the soluble proteins, based on the sequence information. Then, we compared the classification abilities of the two models produced from the different data sets.

In the *in vivo E. coli* expression system, using data set_ME, the statistical models' abilities were estimated using a 5-fold cross validation test (Table II). The accuracies, which signified the proportions of correct prediction, were 77.6 and 71.4%, respectively, for protein expression and solubility in the *in vivo E. coli* expression system. These values were almost identical to those of the models using all features, presented in the 'Materials and Methods' section. Next, we built a statistical model trained using data set_ME, and evaluated its classification ability using data set_SE (Table II). Based on the accuracy (*Acc.*), the classification ability for data set_SE was slightly lower than that for data set_ME. This difference in the ability between the two models is considered to reflect the experimental error that data set_SE includes, because it was much smaller than the experimental error rate inferred from the analysis of data set_ME. Therefore, the two kinds

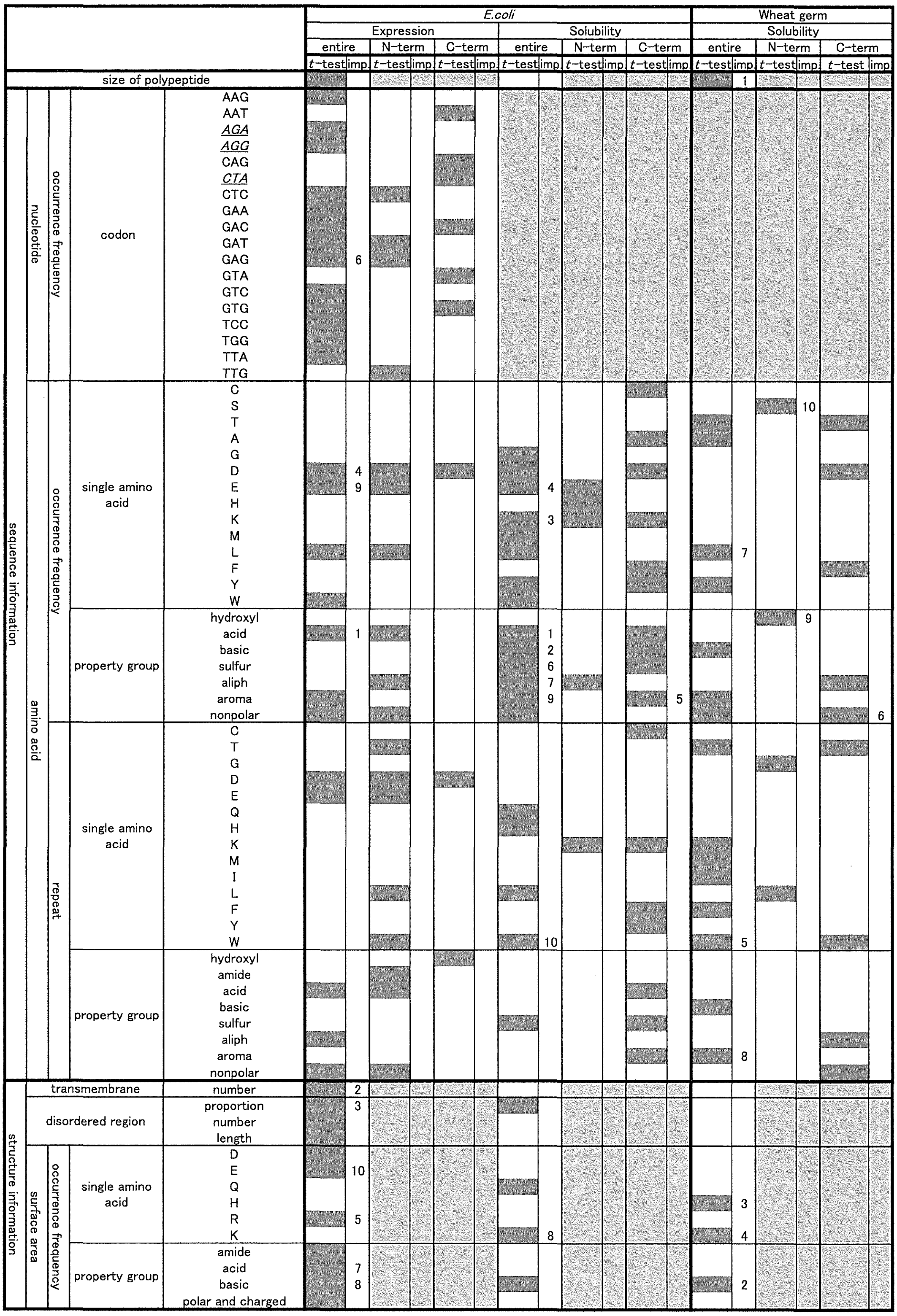

E.coli
Wheat germ
Expression
Solubility
Solubility
entire
N-term
C-term
t-test
imp.
size of polypeptide
1
sequence information
nucleotide
occurrence frequency
codon
AAG
AAT
AGA
AGG
CAG
CTA
CTC
GAA
GAC
GAT
GAG
GTA
GTC
GTG
TCC
TGG
TTA
TTG
6
amino acid
single amino acid
C
S
T
A
G
D
E
H
K
M
L
F
Y
W
4
9
4
3
10
7
property group
hydroxyl
acid
basic
sulfur
aliph
aroma
nonpolar
1
1
2
6
7
9
9
5
6
repeat
single amino acid
C
T
G
D
E
Q
H
K
M
I
L
F
Y
W
10
5
property group
hydroxyl
amide
acid
basic
sulfur
aliph
aroma
nonpolar
8
structure information
transmembrane
number
2
disordered region
proportion
number
length
3
surface area
occurrence frequency
single amino acid
D
E
Q
H
R
K
10
5
8
3
4
property group
amide
acid
basic
polar and charged
7
8
2

Fig. 4 Comparison of features associated with protein expression/solubility in the two expression systems. Only the features with statistically significant differences detected by the Student's *t*-test are listed in this figure. *t*-test shows the results of the Student's *t*-test. Red signifies the features that have a positive effect on protein expression or solubility, and blue shows the features that have a negative influence. White denotes features not found to have a statistically significant difference; grey shows that no statistical test was done. Entire, N-term, and C-term signify features computed using the entire chain, the N-terminal region, and the C-terminal region, respectively. 'imp.' shows the ranking of features that contribute to protein expression/solubility. The ranking was determined for three categories: protein expression in the *in vivo E. coli*, protein solubility in the *in vivo E. coli*, and protein solubility in wheat germ. The rare-frequency codons in the genetic code of *E. coli* are italicized and underlined.

Table II. Classification abilities of protein expression/solubility in the two expression systems.

		Expression			Solubility		
Expression system	Data set	Recall	Precision	*Acc.*	Recall	Precision	*Acc.*
Escherichia coli	Data set_ME	0.807	0.838	0.776	0.673 (0.296)	0.468 (0.429)	0.714 (0.587)
	Data set_SE	0.876	0.702	0.694	0.424 (0.295)	0.551 (0.432)	0.671 (0.610)
Wheat Germ	Data set_MW	–	–	–	0.736 (0.302)	0.853 (0.897)	0.714 (0.537)
	Data set_SW	–	–	–	0.892 (0.294)	0.718 (0.846)	0.682 (0.469)

The prediction results were classified into four categories: TP is the number of true positives, which is defined as the number of correctly predicted positives. Similarly, FP, TN and FT denote the numbers of false positives, which are defined, respectively, as: negatives that were incorrectly predicted as positives, the number of true negatives, which are defined as correctly predicted negatives, and the number of false negatives, which are defined as positives incorrectly predicted as negatives. Recall and Precision were defined as [=TP/(TP + FP)] and [=TP/(TP + FN)], respectively. *Acc.* [=(TP + TN)/(TP + TN + FP + FN)] represents the proportion of correctly identified positives plus negatives. A hyphen shows that no statistical test was done. The figures in parentheses signify the results of Wilkinson and Harrison model.

of statistical models are considered to have comparable classification abilities. A similar tendency was observed for the wheat germ cell-free expression system (Table II).

These results indicate that the characteristics of the two pairs of data sets—data set_ME and data set_SE, and data set_MW and data set_SW—are similar. Consequently, the features selected in the previous section represent the general characteristics of the protein expression/solubility in each expression system. Therefore, these features in each expression system are considered to be the minimal sets of features associated with protein expression/solubility.

The RF model can estimate the importance of features more simply than commonly-used machine learning methods, such as support vector machine (SVM) (*30*). We estimated the 10 important features based on the mean degrees of *Acc.* (Fig. 4). A comparison of the two expression systems revealed that the key features associated with protein solubility are different. The features related to charge occupied the top rank in the *in vivo E. coli* expression system, while they are hardly found in the wheat germ cell-free expression system.

Discussion

We identified a minimal set of features associated with protein expression/solubility in two expression systems, by the application of two statistical analyses. A comparison of the features associated with protein expression/solubility in the *in vivo E. coli* expression system revealed their different influences. In short, the 'structural information' has a strong influence at the protein expression stage, whereas the amino acid 'sequence information' exerts effects at the protein solubility stage (Fig. 4). These observations suggest a mechanism for yielding a soluble protein in the *in vivo E. coli* expression system. Regarding the protein expression stage, increased numbers of disordered regions and transmembrane regions act to prevent protein expression. Experiments with individual proteins have also shown that disordered regions affect protein expression (*31*). In addition, the presence of charged residues on the protein surface has a positive effect on protein expression. These are common characteristics of globular proteins. For this reason, it may be important for a protein to fold into the proper structure at the protein expression stage. In contrast, the amino acid sequence information is important for the solubility stage. The statistical analysis indicated that an abundance of charged residues in the C-terminal region leads increase of protein solubility. In a study of an individual protein, Kato *et al.* (*32*) reported that adding several arginine residues to the C-terminus of BPTI increased its solubility by preventing aggregation. Therefore, it may be important for a protein not to aggregate at the protein solubility stage.

A comparison of the two expression systems revealed two important points. One is that the number of features associated with protein solubility in the wheat germ cell-free expression system is smaller than that in the *in vivo E. coli* expression system (Fig. 4). This observation implies that the wheat germ cell-free expression system is less sensitive to the various sequence and structural features of a protein, corresponding to the fact that the wheat germ cell-free expression system has a higher success rate than the *in vivo E. coli* expression system in generating soluble proteins (Table I). The other is that the key features in the two expression systems are different. In the *in vivo E. coli* expression system, the charge is important, but it has little influence on the solubility in the wheat germ cell-free expression system. The differences between the features in the two expression systems might be related to the translation speed (*33*). In general, the speed is faster in bacteria than in eukaryotes. The charged residues are considered to be important for partial fording in the *in vivo E. coli* expression system.

The minimal sets of features associated with protein expression/solubility in the two expression systems are useful to screen targets in protein expression experiments. When the statistical model that used the minimal set of features identified in this study was compared with Wilkinson's statistical model (*11*) to predict the *in vitro* solubility of a recombinant protein in an *E. coli* expression system, the *Acc.* of our model for data set_SE was 6.1% higher than that of Wilkinson's model.

Supplementary Data

Supplementary Data are available at *JB* Online.

Acknowledgements

We thank the members of the molecular function team at the Computational Biology Research Center (CBRC) and Dr Kuroda at Tokyo University of Agriculture and Technology (TUAT) for helpful discussions and advice.

Funding

The Okawa Foundation for Information and Telecommunications.

Conflict of interest

None declared.

References

1. Clark, E.D.B. (1998) Refolding of recombinant proteins. *Curr. Opin. Biotechnol.* **9**, 157–163
2. Doray, B., Chen, C.D., and Kemper, B. (2001) N-terminal deletions and His-tag fusions dramatically affect expression of cytochrome p450 2C2 in bacteria. *Arch. Biochem. Biophys.* **393**, 143–153
3. Sati, S.P., Singh, S.K., Kumar, N., and Sharma, A. (2002) Extra terminal residues have a profound effect on the folding and solubility of a Plasmodium falciparum sexual stage-specific protein over-expressed in *Escherichia coli. Eur. J. Biochem.* **269**, 5259–5263
4. Kapust, R.B. and Waugh, D.S. (1999) *Escherichia coli* maltose-binding protein is uncommonly effective at promoting the solubility of polypeptides to which it is fused. *Protein Sci.* **8**, 1668–1674
5. Tresaugues, L., Collinet, B., Minard, P., Henckes, G., Aufrere, R., Blondeau, K., Liger, D., Zhou, C.Z., Janin, J., Van Tilbeurgh, H., and Quevillon-Cheruel, S. (2004) Refolding strategies from inclusion bodies in a structural genomics project. *J. Struct. Funct. Genomics* **5**, 195–204
6. Andersen, D.C. and Krummen, L. (2002) Recombinant protein expression for therapeutic applications. *Curr. Opin. Biotechnol.* **13**, 117–123
7. Kramer, G., Kudlicki, W., Hardesty, B., Higgens, S.J., and Hames, B.D. (1999) Cell-free coupled transcription-translation systems from *Escherichia coli.* In *Protein Expression: A Practical Approach* (Higgens, S.J. and Hames, B.D., eds.), pp. 201–223, Oxford University Press, Oxford
8. Clemens, M.M., Prujin, G.J., Higgens, S.J., and Hames, B.D. (1999) Protein synthesis in eukaryotic cell-free systems. In *Protein Expression. A Practical Approach* (Higgens, S.J. and Hames, B.D., eds.), pp. 129–165, Oxford University Press, Oxford
9. Goshima, N., Kawamura, Y., Fukumoto, A., Miura, A., Honma, R., Satoh, R., Wakamatsu, A., Yamamoto, J., Kimura, K., Nishikawa, T., Andoh, T., Iida, Y., Ishikawa, K., Ito, E., Kagawa, N., Kaminaga, C., Kanehori, K., Kawakami, B., Kenmochi, K., Kimura, R., Kobayashi, M., Kuroita, T., Kuwayama, H., Maruyama, Y., Matsuo, K., Minami, K., Mitsubori, M., Mori, M., Morishita, R., Murase, A., Nishikawa, A., Nishikawa, S., Okamoto, T., Sakagami, N., Sakamoto, Y., Sasaki, Y., Seki, T., Sono, S., Sugiyama, A., Sumiya, T., Takayama, T., Takayama, Y., Takeda, H., Togashi, T., Yahata, K., Yamada, H., Yanagisawa, Y., Endo, Y., Imamoto, F., Kisu, Y., Tanaka, S., Isogai, T., Imai, J., Watanabe, S., and Nomura, N. (2008) Human protein factory for converting the transcriptome into an in vitro-expressed proteome. *Nat. Methods* **5**, 1011–1017
10. He, M. (2008) Cell-free protein synthesis: applications in proteomics and biotechnology. *N. Biotechnol.* **25**, 126–132
11. Wilkinson, D.L. and Harrison, R.G. (1991) Predicting the solubility of recombinant proteins in *Escherichia coli. Biotechnology* **9**, 443–448
12. Christendat, D., Yee, A., Dharamsi, A., Kluger, Y., Savchenko, A., Cort, J.R., Booth, V., Mackereth, C.D., Saridakis, V., Ekiel, I., Kozlov, G., Maxwell, K.L., Wu, N., McIntosh, L.P., Gehring, K., Kennedy, M.A., Davidson, A.R., Pai, E. F., Gerstein, M., Edwards, A.M., and Arrowsmith, C.H. (2000) Structural proteomics of an archaeon. *Nat. Struct. Biol.* **7**, 903–909
13. Bertone, P., Kluger, Y., Lan, N., Zheng, D., Christendat, D., Yee, A., Edwards, A.M., Arrowsmith, C.H., Montelione, G.T., and Gerstein, M. (2001) SPINE: an integrated tracking database and data mining approach for identifying feasible targets in high-throughput structural proteomics. *Nucleic Acids Res.* **29**, 2884–2898
14. Goh, C.S., Lan, N., Douglas, S.M., Wu, B., Echols, N., Smith, A., Milburn, D., Montelione, G.T., Zhao, H., and Gerstein, M. (2004) Mining the structural genomics pipeline: identification of protein properties that affect high-throughput experimental analysis. *J. Mol. Biol.* **336**, 115–130
15. Luan, C.H., Qiu, S., Finley, J.B., Carson, M., Gray, R.J., Huang, W., Johnson, D., Tsao, J., Reboul, J., Vaglio, P., Hill, D.E., Vidal, M., Delucas, L.J., and Luo, M. (2004) High-throughput expression of *C. elegans* proteins. *Genome Res.* **14**, 2102–2010
16. Idicula-Thomas, S. and Balaji, P.V. (2005) Understanding the relationship between the primary structure of proteins and its propensity to be soluble on overexpression in *Escherichia coli. Protein Sci.* **14**, 582–592
17. Niwa, T., Ying, B.W., Saito, K., Jin, W., Takada, S., Ueda, T., and Taguchi, H. (2009) Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of *Escherichia coli* proteins. *Proc. Natl Acad. Sci. USA* **106**, 4201–4206
18. Maruyama, Y., Wakamatsu, A., Kawamura, Y., Kimura, K., Yamamoto, J., Nishikawa, T., Kisu, Y., Sugano, S., Goshima, N., Isogai, T., and Nomura, N. (2009) Human Gene and Protein Database (HGPD): a novel database presenting a large quantity of experiment-based results in human proteomics. *Nucleic Acids Res.* **37**, 762–766
19. Li, W. and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659
20. Rost, B. (1996) PHD: predicting one-dimensional protein structure by profile-based neural networks. *Methods Enzymol.* **266**, 525–539
21. Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305**, 567–580
22. Hirose, S., Shimizu, K., Kanai, S., Kuroda, Y., and Noguchi, T. (2007) POODLE-L: a two-level SVM prediction system for reliably predicting long disordered regions. *Bioinformatics* **23**, 2046–2053
23. Ahmad, S., Gromiha, M.M., and Sarai, A. (2003) RVP-net: online prediction of real valued accessible surface area of proteins from single sequences. *Bioinformatics* **19**, 1849–1851

24. Liaw, A. and Wiener, M. (2002) Classification and Regression by randomForest. *R News* **2**, 18–22
25. Makrides, S.C. (1996) Strategies for achieving high-level expression of genes in *Escherichia coli*. *Microbiol. Rev.* **60**, 512–538
26. Drummond, D.A. and Wilke, C.O. (2008) Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* **134**, 341–352
27. Lorimer, D., Raymond, A., Walchli, J., Mixon, M., Barrow, A., Wallace, E., Grice, R., Burgin, A., and Stewart, L. (2009) Gene composer: database software for protein construct design, codon engineering, and gene synthesis. *BMC Biotechnol.* **9**, 36
28. Welch, M., Govindarajan, S., Ness, J.E., Villalobos, A., Gurney, A., Minshull, J., and Gustafsson, C. (2009) Design parameters to control synthetic gene expression in *Escherichia coli*. *PLoS One* **4**, e7002
29. Cheng, J., Sweredoski, M., and Baldi, P. (2006) DOMpro: protein domain prediction using profiles, secondary structure relative solvent accessibility, and recursive neural network. *Data Min. Knowl. Disc.* **13**, 1–10
30. Breinman, L. (2001) Random forests. *Mach. Learn.* **45**, 5–32
31. Quevillon-Cheruel, S., Leulliot, N., Gentils, L., van Tilbeurgh, H., and Poupon, A. (2007) Production and crystallization of protein domains: how useful are disorder predictions? *Curr. Protein Pept. Sci.* **8**, 151–160
32. Kato, A., Maki, K., Ebina, T., Kuwajima, K., Soda, K., and Kuroda, Y. (2007) Mutational analysis of protein solubility enhancement using short peptide tags. *Biopolymers* **85**, 12–18
33. Siller, E., DeZwaan, D.C., Anderson, J.F., Freeman, B.C., and Barral, J.M. (2010) Slowing bacterial translation speed enhances eukaryotic protein folding efficiency. *J. Mol. Biol.* **396**, 1310–1318

DOI: 10.1002/chem.201101058

Synthesis of Skeletal Analogues of Saxitoxin Derivatives and Evaluation of Their Inhibitory Activity on Sodium Ion Channels Na$_V$1.4 and Na$_V$1.5**

Ryoko Shinohara,[a] Takafumi Akimoto,[a] Osamu Iwamoto,[a] Takatsugu Hirokawa,[b] Mari Yotsu-Yamashita,[c] Kaoru Yamaoka,[d] and Kazuo Nagasawa*[a]

Abstract: Skeletal analogues of saxitoxin (STX) that possess a fused-type tricyclic ring system, designated FD-STX, were synthesized as candidate sodium ion channel modulators. Three kinds of FD-STX derivatives **4a–c** with different substitution at C13 were synthesized, and their inhibitory activity on sodium ion channels was examined by means of cell-based assay. (−)-FD-STX (**4a**) and (−)-FD-dcSTX (**4b**), which showed moderate inhibitory activity, were further evaluated by the use of the patch-clamp method in cells that expressed Na$_V$1.4 (a tetrodotoxin-sensitive sodium channel subtype) and Na$_V$1.5 (a tetrodotoxin-resistant sodium channel subtype). These compounds showed moderate inhibitory activity towards Na$_V$1.4, and weaker inhibitory activity towards Na$_V$1.5. Uniquely, however, the inhibition of Na$_V$1.5 by (−)-FD-dcSTX (**4b**) was "irreversible".

Keywords: inhibitors · ion channels · patch-clamp method · saxitoxins · sodium · structure–activity relationships

Introduction

Voltage-gated sodium channels (Na$_V$Ch) are transmembrane proteins that provide inward current carried by sodium ions, and they contribute to the control of membrane excitability, as well as the propagation of action potentials along axons.[1] Sodium channels within neurons are composed of a single α subunit, which forms the voltage-sensing and ion-selective pore, and one or more auxiliary β subunits, which are proposed to serve a number of functions, including modulation of α-subunit function and targeting/anchoring the channels at specific sites in the plasma membrane.[2] To date, nine genes that encode α subunits (Na$_V$1.1–Na$_V$1.9) and four genes that encode β subunits have been identified. The α subunits can be classified into two groups on the basis of their sensitivity to tetrodotoxin (TTX), the pufferfish toxin. There are six TTX-sensitive (TTX-s) α subunits (i.e., Na$_V$1.1–Na$_V$1.4, Na$_V$1.6, and Na$_V$1.7) and three TTX-resistant (TTX-r) α subunits (i.e., Na$_V$1.5, Na$_V$1.8, and Na$_V$1.9), which are only blocked by high concentrations of TTX.[3,4] Recently, a related protein (Na$_X$X) has been recognized as a tenth member of the group.[5] These subtypes have quite similar structures, but the expression of the α subunits is strongly cell-type- and tissue-specific. Therefore, each of these subtypes is believed to have unique properties. Indeed, mutation studies have thrown some light on the functions of these channels. For example, mutation of Na$_V$1.4 is associated with primary periodic paralysis,[6] whereas mutation of Na$_V$1.5 leads to cardiovascular syndromes.[7] Consequently, these channels are considered to be good candidates for new drug targets. In this context, Na$_V$Ch-subtype-selective small-molecular modulators from natural sources have been explored, as well as synthetic compounds. Synthetic studies have led to some subtype-selective compounds.[8] For example, A-803467 is an Na$_V$1.8-selective blocker developed by an Abbott research group[8c,9] for possible application in pain control. However, more selective modulators with a variety of action mechanisms are still of great interest.

Saxitoxin (STX) (**1a**) is a naturally occurring shellfish neurotoxin that blocks ion influx to TTX-s Na$_V$Chs.[10] STX (**1a**) is an indispensable tool for studying Na$_V$Ch-related electrophysiology, and its derivatives are also promising candidates for Na$_V$Ch-subtype-selective modulators. STX (**1a**) and its analogues are believed to bind to the P-loop region of the ion-selective filter in Na$_V$Ch in a similar manner to

[a] R. Shinohara, T. Akimoto, Dr. O. Iwamoto, Prof. Dr. K. Nagasawa
Tokyo University of Agriculture and Technology
Department of Biotechnology and Life Science
2-24-16 Naka-cho, Koganei, Tokyo 184-8588 (Japan)
Fax: (+81) 42-388-7295
E-mail: knaga@cc.tuat.ac.jp

[b] Dr. T. Hirokawa
Computational Biology Research Center
National Institute of Advanced Industrial
Science and Technology, 2-4-7 Aomi, Koto-ku
Tokyo 135-0064 (Japan)

[c] Prof. Dr. M. Yotsu-Yamashita
Tohoku University, Graduate of Agricultural Science
1-1 Tsutsumidori-Amamiyamachi, Aoda-ku
Sendai 981-8555 (Japan)

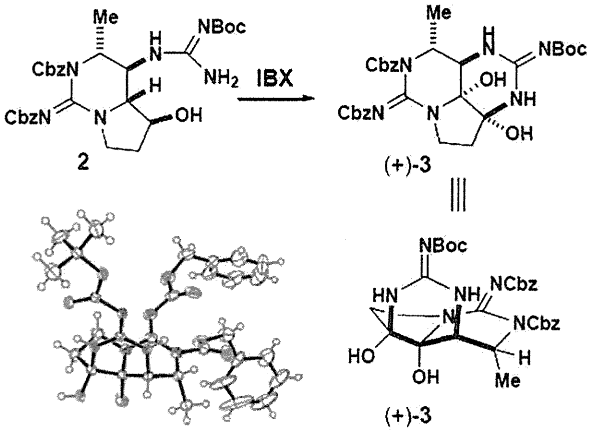
[d] Prof. Dr. K. Yamaoka
Hiroshima International University
555-36 Kurose-Gakuendai, Higashi-Hiroshima
Hiroshima 739-2695 (Japan)

[**] Na$_V$Ch denotes voltage-gated sodium channels. Na$_V$1.4 is a tetrodotoxin-sensitive sodium channel subtype; Na$_V$1.5 is a tetrodotoxin-resistant sodium channel subtype.

Supporting information for this article is available on the WWW under http://dx.doi.org/10.1002/chem.201101058.

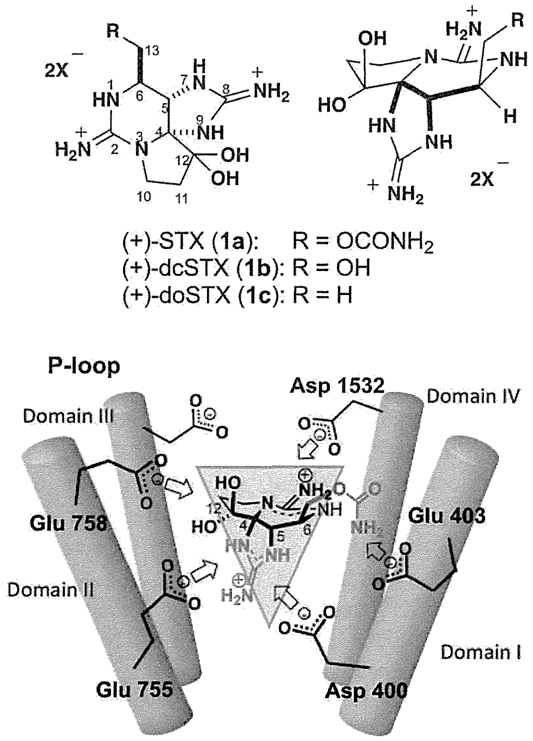
TTX (i.e., in a reversible manner).[11] Recently, a binding model of STX with the P-loop domain was proposed based on molecular docking studies, and it was suggested that the two guanidinium groups, the carbamoyl group at C13 and the hydrated ketone at C12, have critical roles in the binding.[12] On the other hand, most naturally occurring STX analogues isolated to date are functionalized at C11, C12, C13, N1, and N7, and these modifications have a significant influence on the Na_VCh-inhibitory activity.[13] Thus, structure–activity relationship (SAR) studies of STX have been focused on those positions,[14] especially C13, with the aim of developing Na_VCh-subtype-selective modulators.

We have recently reported a total synthesis of STX (**1a**; see Scheme 1) and some of its naturally occurring derivatives.[15] In the course of our synthetic studies of these compounds, we unexpectedly obtained a fused-type tricyclic compound (+)-**3** during the oxidation of **2** with 2-iodoxybenzoic acid (IBX; Scheme 2).[15a] An X-ray analysis showed compound **3** to have a conical shape similar to that of STX (see Scheme 1), and compound **3** was expected to be able to access the channel conical vestibule in a similar manner to STX. Thus, this fused-type tricyclic skeleton can be regarded as a new skeletal analogue of STX, and we named it FD-STX. The FD-STX structure has almost the same functional groups as STX at positions that interact with Na_VCh, although the angles and/or directions of functional groups are slightly different from those of STX.[16] We considered that these small structural differences of FD-STX compared with STX might represent an opportunity to acquire new subtype selectivity and/or might result in a distinctive interaction mode of FD-STX with Na_VChs.



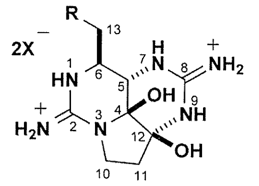
Scheme 2. Synthesis of fused-type tricyclic compound (+)-**3**.

In this report, we describe the synthesis of FD-STX derivatives (−)-**4a–c** functionalized at C13, which correspond to the skeletal analogues of (+)-STX (**1a**), (+)-dcSTX (**1b**), and (+)-doSTX (**1c**), respectively (Scheme 3). The Na_VCh-inhibitory activity of these FD-STX derivatives was initially examined by cell-based assay, and (−)-FD-STX (**4a**) and (−)-FD-dcSTX (**4b**) were further evaluated by means of the patch-clamp method in cells that express Na_V1.4 or Na_V1.5. The results are of particular interest because no SAR studies that focus on the STX skeleton itself have been reported so far.



(+)-STX (**1a**): R = $OCONH_2$
(+)-dcSTX (**1b**): R = OH
(+)-doSTX (**1c**): R = H

Scheme 1. Structures of saxitoxin (STX) and its derivatives, and a model of their binding to the sodium ion channel.[12c]

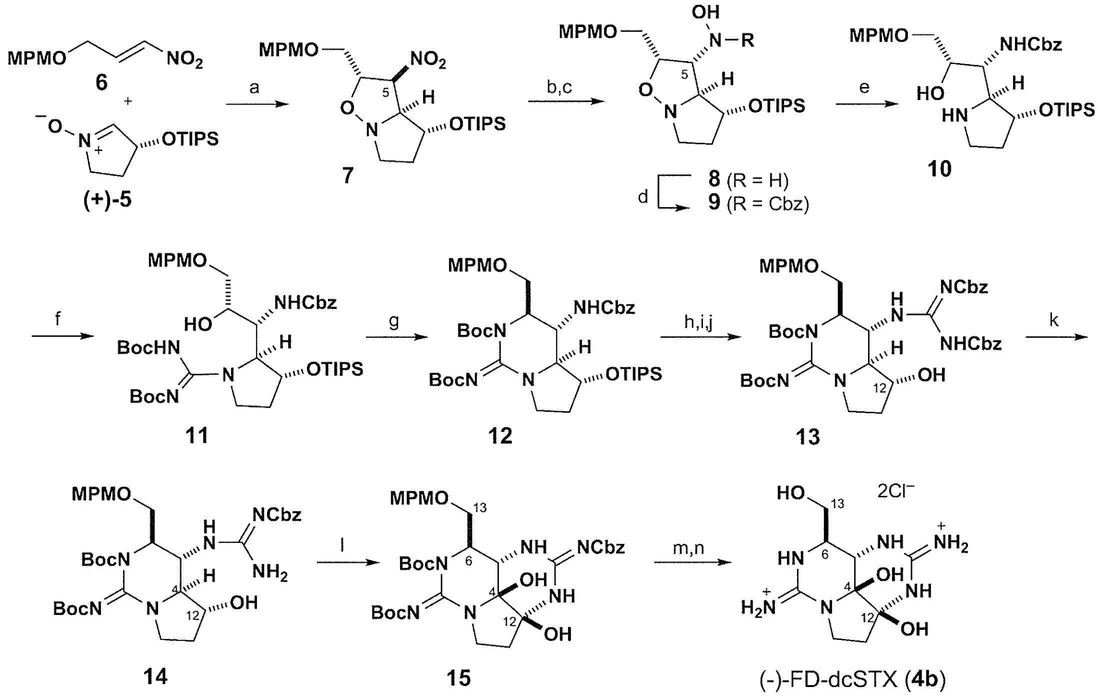


(−)-FD-STX (**4a**): R = $OCONH_2$
(−)-FD-dcSTX (**4b**): R = OH
(−)-FD-doSTX (**4c**): R = H

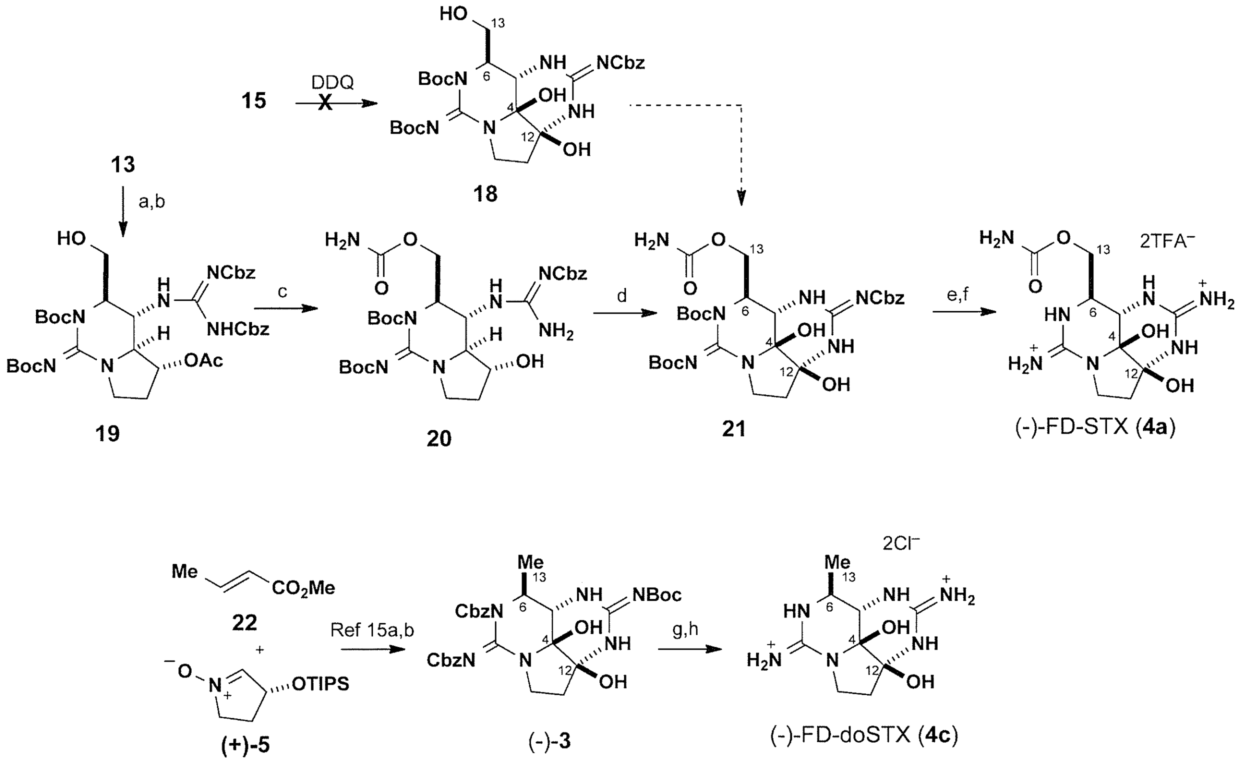
Scheme 3. Structures of (−)-FD-STXs **4a–c**, skeletal analogues of (+)-STXs **1a–c**.

Results and Discussion

Synthesis of (−)-FD-STX and its derivatives (4a–c): Scheme 4 and Scheme 5 illustrate the synthesis of **4a–c**. We first focused on the synthesis of (−)-FD-dcSTX (**4b**) (Scheme 4).[15] 1,3-Dipolar cycloaddition reaction between chiral nitrone **5** and nitro alkene **6** at 40 °C gave the *endo* adduct of isoxazolidine **7** as a single diastereomer. The stereochemistry at C5 was epimerized with 1,8-diazabicyclo-[5.4.0]undec-7-ene (DBU) in dichloromethane at 0 °C (95:5), and selective reduction of the nitro group was carried out with zinc powder in acetic acid to give hydroxylamine **8**



Scheme 4. Synthesis of (−)-FD-dcSTX (**4b**): a) 40 °C; b) DBU, CH_2Cl_2, −40 °C; c) AcOH, then Zn, −40 °C, 72 % (3 steps); d) CbzCl, K_2CO_3, THF/H_2O = 3:1, 0 °C, 83 %; e) NaOAc, $TiCl_3$/HCl, Zn, CH_2Cl_2/MeOH = 2:1, −50 °C; f) BocN = C(SMe)NHBoc (**16**), $HgCl_2$, Et_3N, DMF, 82 % (2 steps); g) $ClCH_2SO_2Cl$, *i*Pr_2EtN, CH_2Cl_2, 87 %; h) TBAF, THF, 0 °C, 93 %; i) H_2, $Pd(OH)_2/C$, MeOH; j) CbzN = C(SMe)NHCbz (**17**), $HgCl_2$, Et_3N, DMF, 85 % (2 steps); k) NaOMe, THF/MeOH, 0 °C, 65 %; l) IBX, DMSO, 70 °C, 24 %; m) H_2, $Pd(OH)_2/C$, MeOH; n) 3 N HCl, 78 % (2 steps).



Scheme 5. Synthesis of (−)-FD-STX (**4a**): a) Ac_2O, pyridine (py), 50 °C, 96 %; b) DDQ, CH_2Cl_2, H_2O, 94 %; c) trichloroacetyl isocyanate, CH_2Cl_2, then K_2CO_3, MeOH, 34 %; d) IBX, DMSO, 70 °C, 20 %; e) H_2, $Pd(OH)_2/C$, MeOH; f) TFA, CH_2Cl_2, 82 % (2 steps). Synthesis of (−)-FD-doSTX (**4c**): g) TFA, CH_2Cl_2; h) H_2, Pd/C, MeOH, 92 % (2 steps).