Structural Determinants in Protein Folding: A Single Conserved Hydrophobic Residue Determines Folding of EGF Domains

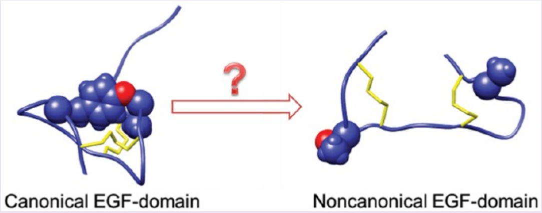
A. S. Angie Ng[†,∥] and R. Manjunatha Kini*[,†,‡,§]

[†]Department of Biological Sciences, Faculty of Science, National University of Singapore, Singapore 117543, Singapore

[‡]Department of Biochemistry, Medical College of Virginia, Virginia Commonwealth University, Richmond, Virginia 23298-0614, United States

[§]School of Pharmacy and Medical Sciences, University of South Australia, Adelaide, South Australia 5000, Australia

S Supporting Information



ABSTRACT: The epidermal growth factor (EGF) domain is evolutionarily conserved despite hypervariability in amino acid sequences. They fold into a three-looped conformation with a disulfide pairing of C_1-C_3, C_2-C_4 ,and C_5-C_6. To elucidate the structural determinants that dictate the EGF fold, we selected the fourth and fifth EGF domains of thrombomodulin (TM) as models; the former domain folds into the canonical conformation, while the latter domain folds with alternate disulfide pairing of C_1-C_2, C_3-C_4, and C_5-C_6. Since their third disulfide (C_5-C_6) is conserved, we examined the folding tendencies of synthetic peptides corresponding to truncated domain four (t-TMEGF4) and five (t-TMEGF5), encompassing the segment C_1 to C_4. These peptides fold into their respective disulfide isoforms indicating that they contain all the required structural determinants. On the basis of the folding tendencies of these peptides in the absence and presence of 6 M Gn·HCl or 0.5 M NaCl, we determined that hydrophobic interactions are needed for the canonical EGF fold but not for the noncanonical fold. Sequence alignment of extant EGF domains and examination of their three-dimensional structures allowed us to identify a highly conserved hydrophobic residue in intercysteine loop 3 as the key contributor, which nucleates the hydrophobic core and acts as the lynch pin. When this hydrophobic residue (Tyr25) was substituted with a more hydrophilic Thr, the hydrophobic interactions were disrupted, and t-TMEGF4-Y25T folds similar to t-TMEGF5. Taken together, our results for the first time demonstrate that a single conserved hydrophobic residue acts as the key determinant in the folding of EGF domains.

Protein folding remains one of the most intriguing question of structural biology. Although the Anfinsen's thermodynamic hypothesis[1,2] provided an apparent answer to some parts of the question, the mechanistic details of how it works remain elusive. The enigma and ambiguity in the protein folding problem are further enhanced by the structural conservation of protein domains despite extreme hypervariability in amino acid sequences. The epidermal growth factor (EGF) domain, a functionally diverse building block for extracellular proteins, is an excellent example of extreme sequence hypervariability in a structurally conserved protein domain.[3] This domain has 30–40 amino acid residues with six conserved cysteines. It forms a three-looped structure made up of a central two-stranded β-sheet followed by a loop to a short C-terminal two-stranded sheet. This highly conserved fold is stabilized by three disulfide bonds formed between the first and third (C_1-C_3), second and fourth (C_2-C_4), and fifth and sixth (C_5-C_6) cysteine residues. Although the functional diversity of EGF domains could be explained by the hypervariability of amino acid sequences in their intercysteine regions, the structural determinants that conserve their scaffold structure are not known.

Thrombomodulin (TM) is a transmembrane glycoprotein expressed on the surface of vascular endothelial cells, and it acts as a cofactor in switching thrombin from a procoagulant to an anticoagulant enzyme.[4,5] It has six EGF domains, with the fourth and fifth EGF domains (TMEGF4 and TMEGF5) constituting the smallest cofactor-active fragment;[6] while TMEGF4 contributes to the cofactor activity of TM, TMEGF5 is essential for anchoring TMEGF4 to thrombin. Interestingly, TMEGF4 folds into the canonical EGF domain structure defined by C_1-C_3, C_2-C_4, and C_5-C_6 disulfide-connectivity[7,8] (Figure 1a), while TMEGF5 folds into an atypical structure defined by C_1-C_2, C_3-C_4, and C_5-C_6 disulfide-connectivity[8,9] (Figure 1b). Thus, in TMEGF5, the central two-stranded β-sheet of canonical EGF domains is absent, with its N- and C-termini being closer together than in the canonical structure. The functional significance of this noncanonical structure was examined by Hunter and Komives,[10] where different disulfide-bonded isomers of TMEGF5 were tested for their thrombin-binding affinities. Their studies showed that the native C_1-C_2, C_3-C_4, and C_5-C_6 isoform of TMEGF5 has a higher affinity for thrombin compared to its corresponding C_1-C_3, C_2-C_4, and C_5-C_6 isoform. This suggests that the noncanonical structure of

Received: August 23, 2012
Accepted: October 5, 2012
Published: October 24, 2012

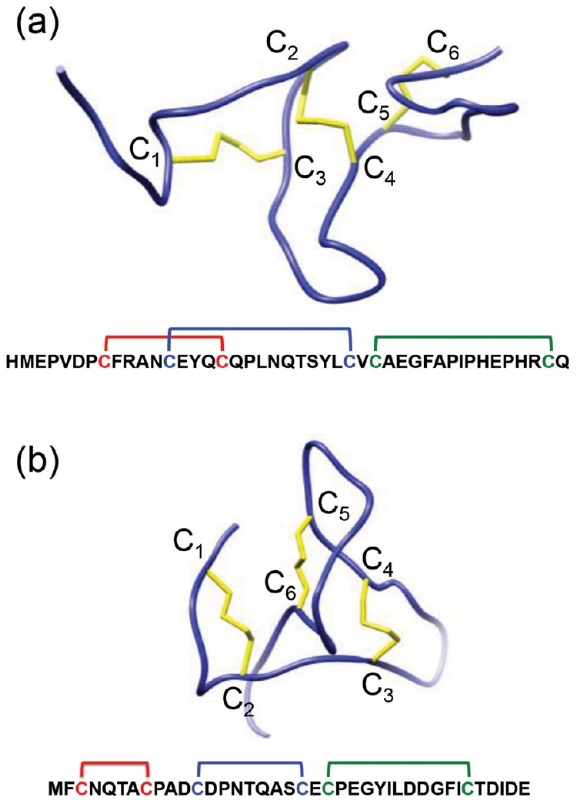

(a)
C2
C6
C5
C1
C3
C4
HMEPVDPCFRANCEYQCQPLNQTSYLCVCAEGFAPIPHEPHRCQ
(b)
C5
C1
C4
C6
C3
C2
MFCNQTACPADCDPNTQASCECPEGYILDDGFICTDIDE

Figure 1. Disulfide pairing in TMEGF4 and TMEGF5 domains. Backbone structure of (a) TMEGF4 and (b) TMEGF5 [PDB: 1DQB]. Cysteine residues are labeled as C_1, C_2, C_3, C_4, C_5, and C_6, and disulfide linkages are indicated in yellow. TMEGF4 shows canonical disulfide-connectivity of C_1-C_3, C_2-C_4, and C_5-C_6. In contrast, TMEGF5 shows noncanonical disulfide-connectivity of C_1-C_2, C_3-C_4, and C_5-C_6.

TMEGF5 has a functional significance in the anticoagulant activity of TM and is thus evolutionarily selected.

In this study, we have used TMEGF4 and TMEGF5 as contrasting models to elucidate the structural determinants that define the folding of EGF domains. By determining the folding tendencies of various synthetic peptides designed based on these two domains and by comparing amino acid sequences and three-dimensional structures of EGF domains, we identified the importance of the hydrophobic core for the conservation of the canonical EGF fold. A single highly conserved hydrophobic residue in the penultimate position in intercysteine loop 3 plays the key role in the formation of this core. This residue and the hydrophobic core are not conserved in TMEGF5, thus defining its noncanonical fold. We have shown that the replacement of this hydrophobic residue results in the disruption of the hydrophobic core leading to altered disulfide pairing. Thus, for the first time, we describe the role of the hydrophobic residue in intercysteine loop 3 as the key structural determinant that defines this common, multifunctional domain.

■ RESULTS

Synthesis of Truncated TMEGF4 and TMEGF5 Structural Isoforms. Although TMEGF4 and TMEGF5 show distinct folding due to altered disulfide pairing, the difference in disulfide-connectivity is restricted to the first two disulfide bonds in their N-terminal segments (encompassing C_1 to C_4). Thus, it was of interest to determine whether the structural determinants are located within this segment or the C-terminal segment (encompassing C_5 to C_6). Therefore, we synthesized truncated versions of TMEGF4 and TMEGF5 (t-TMEGF4 and t-TMEGF5) without the segment encompassing C_5 to C_6 for in vitro oxidative folding studies. We synthesized all three structural isoforms (C_1-C_3, C_2-C_4; C_1-C_2, C_3-C_4; and C_1-C_4, C_2-C_3) of t-TMEGF4 and t-TMEGF5 using regioselective incorporation of cysteine residues (Supplementary Figure S1). The observed average masses of these isoforms corresponded well with the theoretical (fully oxidized) average mass of 3284.7 and 2244.5 Da for t-TMEGF4 and t-TMEGF5, respectively (Supplementary Table S1). The retention volume of each individual structural isoform was determined by reversed-phase HPLC (RP-HPLC).

Folding Tendencies of t-TMEGF4 and t-TMEGF5. To determine the folding tendencies of t-TMEGF4 and t-TMEGF5, fully reduced peptides were placed in high pH (pH 8.0) buffer. We used two oxidative folding conditions: (a) air oxidation, which makes use of atmospheric oxygen, where the process goes through a series of free radical intermediates;[11] and (b) redox system, which involves the use of reduced/oxidized glutathione at a ratio of 2:1. These compounds catalyze disulfide exchange reactions resulting in the most thermodynamically favorable status of the cysteine residues.[12]

Air oxidation-mediated folding of t-TMEGF4 was monitored by the Ellman's test, and the reaction was deemed complete by 98 h. Structural isoforms obtained from the reaction were resolved by RP-HPLC, and as expected, three monomeric isoforms were obtained (Figure 2a). Each folding isoform was identified by comparing their retention volumes with those of regioselectively synthesized structural isoforms. The relative proportions of each isoform were calculated based on the area of their respective peaks (Table 1a). Similarly, redox reagent-mediated folding of t-TMEGF4 also yielded three monomeric isoforms (Figure 2a; Table 1a). Results from both experiments showed that t-TMEGF4 has a folding preference toward the C_1-C_3, C_2-C_4 (native) isoform (~68–69%) (Figure 3a).

For t-TMEGF5, air oxidation-mediated folding of the reduced peptide was deemed complete by 72 h. Folding of t-TMEGF5 in the redox buffer system was also performed, and in both cases, three monomeric isoforms were obtained (Figure 2b). t-TMEGF5 also has a folding preference toward its native C_1-C_2, C_3-C_4 isoform (~60%) (Table 1b; Figure 3f).

Together, these results demonstrate that fully reduced t-TMEGF4 and t-TMEGF5 peptides preferentially fold into their native structural isoforms even without their C-terminal segment encompassing C_5 to C_6. This suggests the existence of structural determinants required for native disulfide pairing and folding within the N-terminal segments and the C-terminal segments do not play a major role in dictating the disulfide-connectivity of the first two disulfide bonds. In addition, results from the redox reagent-mediated folding experiments suggest that the respective native isoforms of both domains are the most thermodynamically stable among the three possible structural isoforms.

Effect of Denaturant on Folding. To determine the role of side-chain interactions in dictating the folding tendency, 6 M guanidine hydrochloride (Gn·HCl) was included in the oxidative folding buffer. Air oxidation of t-TMEGF4 in the presence of 6 M Gn·HCl was completed by 72 h. Analysis of air oxidation and redox reagent-mediated folding products by RP-HPLC revealed that all three monomeric isoforms were

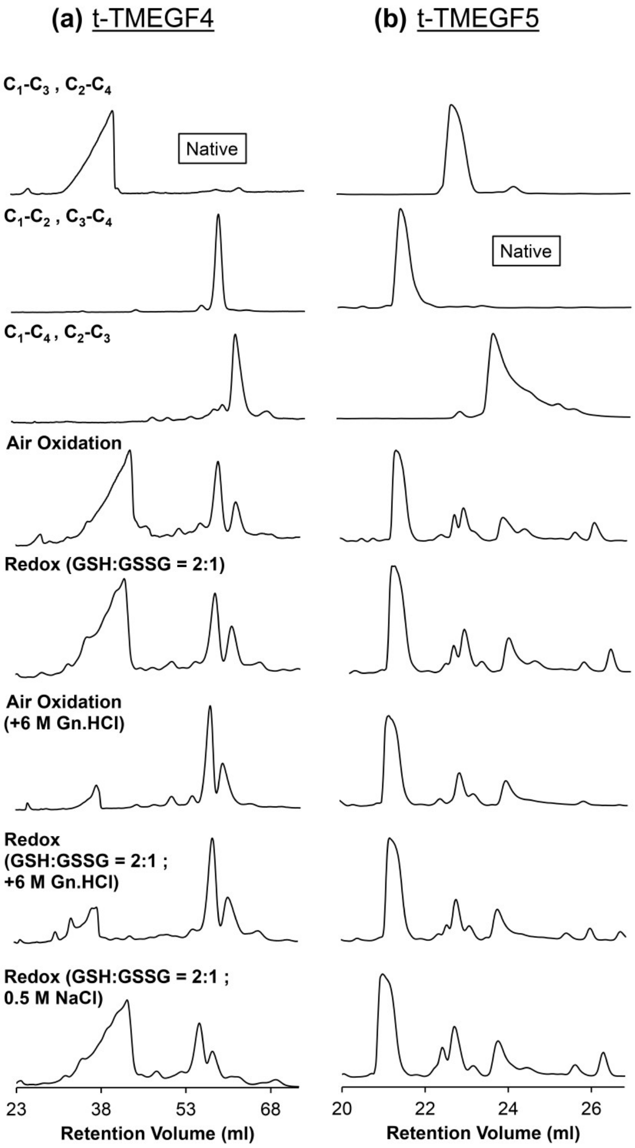


Figure 2. Oxidative folding of t-TMEGF4 and t-TMEGF5 peptides. Structural isoforms obtained from the oxidative folding reactions were separated by chromatography and identified by comparing their respective retention volume with that of regioselectively synthesized isoforms (top three elution profiles; native isoforms are identified). The relative proportion of each isoform was then assessed by measuring its corresponding peak area. The folding conditions used for (a) t-TMEGF4 and (b) t-TMEGF5 peptides were air oxidation and redox reagent-mediated oxidation, in the absence or presence of additives (6 M Gn·HCl or 0.5 M NaCl).

obtained even in the presence of denaturant (Figure 2a). Quantification based on peak areas revealed that the C_1-C_2, C_3-C_4 isoform (~47–53%) was preferred over its native C_1-C_3, C_2-C_4 isoform, when folded in the presence of denaturant (Table 1a; Figure 3b,c,d). Folding of t-TMEGF5 in the presence of 6 M Gn·HCl was also performed via air oxidation and redox reagent-mediated oxidation. For air oxidation, the reaction was completed by 48 h. In both cases, three structural isoforms were obtained (Figure 2b). Quantification of structural isoform proportions revealed that the folding tendency of t-TMEGF5 was not affected by the presence of 6 M Gn·HCl. t-TMEGF5 still showed a folding preference toward its native (C_1-C_2, C_3-C_4) isoform (Table 1b; Figure 3g,h,i).

These results suggest that side-chain interactions are crucial for the formation of the canonical C_1-C_3, C_2-C_4 isoform of t-TMEGF4. When these interactions are neutralized the folding reverts to the C_1-C_2, C_3-C_4 isoform. On the contrary, side-chain interactions do not seem to be crucial for the formation of the native C_1-C_2, C_3-C_4 isoform of t-TMEGF5.

Effect of High Salt Concentration on Folding. To ascertain whether the side-chain interactions that are important for the native C_1-C_3, C_2-C_4 fold of t-TMEGF4 are hydrophobic or electrostatic in nature, we examined the oxidative folding of t-TMEGF4 and t-TMEGF5 in the presence of high salt concentration. To this end, 0.5 M NaCl was included in the redox oxidative folding buffer to disrupt/mask any possible electrostatic forces[13] and to increase the hydrophobic effect.[14]

In the presence of 0.5 M NaCl, there was a significant increase in the formation of the native C_1-C_3, C_2-C_4 isoform in the case of t-TMEGF4 (~75%) (Figures 2a and 3e; Table 1a). As high salt concentration disrupts electrostatic interactions, the unaltered preferential folding of t-TMEGF4 into its native C_1-C_3, C_2-C_4 isoform indicates that electrostatic interactions do not contribute to folding. Further increase in the C_1-C_3, C_2-C_4 isoform is due to the enhanced hydrophobic effect in the presence of 0.5 M NaCl.

In contrast, while t-TMEGF5 still folds predominantly into its native C_1-C_2, C_3-C_4 isoform, there was a significant decrease in its proportion in the presence of 0.5 M NaCl (Figures 2b and 3j; Table 1b). This decrease was accompanied by a concomitant increase in the C_1-C_3, C_2-C_4 isoform. This was probably due to salt-induced increase in the hydrophobic effect.

Together, these results showed that hydrophobic interactions are the dominant force that drives the C_1-C_3, C_2-C_4 fold of the canonical EGF domains.

Identification of the Key Hydrophobic Structural Determinant. To identify potential hydrophobic residues, we aligned t-TMEGF4 and other canonical EGF domains from various proteins whose three-dimensional structures have been solved. Interestingly, only one hydrophobic residue located in the penultimate position in intercysteine loop 3 is conserved in all EGF domains (Figure 4a; Supplementary Figure S2). This hydrophobic residue is also present in TMEGF4 of other organisms but is substituted by less hydrophobic residue in TMEGF5 (Figure 4b). To further understand the role of this hydrophobic residue, the three-dimensional structures of these EGF domains were examined. The conserved residue makes hydrophobic contact with amino acid residues located within the first intercysteine loop to form a hydrophobic core in all EGF domains examined (Supplementary Figure S3). Thus, this conserved residue most likely nucleates the hydrophobic core and acts as the lynch pin.

In TMEGF4, the conserved hydrophobic residue Tyr25 is in close contact with Ala11 of the first intercysteine loop (Figure 5a). On the contrary, the amino acids residues at their equivalent positions (Figure 5b) in TMEGF5 (Thr50 and Ala62) are located on opposite ends and are not in contact (Figure 5c). To test the role of Tyr25, we substituted it with a more hydrophilic Thr residue, which has the hydroxyl group but not the hydrophobic aromatic ring.

Air oxidation-mediated and redox reagent-mediated folding of reduced t-TMEGF4-Y25T was performed over 72 h (as

Table 1. Percentages of Structural Isoforms Obtained from Oxidative Folding of t-TMEGF4, t-TMEGF5 and t-TMEGF4-Y25T in Various Conditions

	proportion of structural isoform (%)				
	TB[a]		TB[a] + 6 M Gn·HCl		TB[a] + 0.5 M NaCl
structural isoform	air oxidation	redox buffer system	air oxidation	redox buffer system	redox buffer system
		(a) t-TMEGF4			
C_1-C_3, C_2-C_4 (native)	67.53 ± 0.69	69.08 ± 0.57	18.48 ± 1.15	31.31 ± 0.98	74.66 ± 0.87
C_1-C_2, C_3-C_4	21.13 ± 0.67	18.96 ± 0.57	52.72 ± 0.94	47.28 ± 0.32	17.18 ± 0.45
C_1-C_4, C_2-C_3	11.34 ± 0.63	11.96 ± 0.27	28.80 ± 0.82	21.42 ± 0.71	8.17 ± 0.49
		(b) t-TMEGF5			
C_1-C_3, C_2-C_4	20.57 ± 0.34	20.36 ± 0.11	17.69 ± 0.35	19.11 ± 0.15	23.32 ± 0.43
C_1-C_2, C_3-C_4 (native)	60.40 ± 0.64	60.67 ± 1.07	61.80 ± 0.91	61.46 ± 0.77	54.40 ± 1.78
C_1-C_4, C_2-C_3	19.03 ± 0.30	18.97 ± 0.98	20.52 ± 0.67	19.43 ± 0.63	22.28 ± 1.98
		(c) t-TMEGF4-Y25T			
C_1-C_3, C_2-C_4 (native)	22.78 ± 0.32	22.27 ± 0.42	8.16 ± 0.33	17.26 ± 0.45	ND[b]
C_1-C_2, C_3-C_4	42.20 ± 0.36	42.25 ± 0.70	54.08 ± 0.52	45.92 ± 0.36	ND[b]
C_1-C_4, C_2-C_3	35.02 ± 0.06	35.48 ± 0.78	37.76 ± 0.19	36.83 ± 0.18	ND[b]

[a] 0.1 M Tris-HCl, pH 8.0. [b] Not determined.

judged by the Ellman's test) and 48 h, respectively. Instead of folding into the canonical C_1-C_3, C_2-C_4 isoform, t-TMEGF4-Y25T displayed a folding preference toward the noncanonical isoform (C_1-C_2, C_3-C_4; ~42%) (Figure 6; Table 1c). The folding propensity of t-TMEGF4-Y25T in the absence of denaturant is similar to that of t-TMEGF4 in the presence of 6 M Gn·HCl (Figure 7), thus suggesting that the key side-chain interactions disrupted is that of Tyr25. Further, the proportion of the C_1-C_3, C_2-C_4 isoform obtained from the folding of t-TMEGF4-Y25T and t-TMEGF5 (which lacks the hydrophobic residue) under normal oxidative conditions are similar.

The folding tendency of t-TMEGF4-Y25T was unaltered despite the presence of 6 M Gn·HCl in the folding buffer (Figure 6; Table 1c). Similar to t-TMEGF5, such hydrophobic interactions are not found in t-TMEGF4-Y25T, and hence, the presence of denaturant did not change the folding tendency of this EGF domain.

These observations *en masse* strongly support the importance of the conserved hydrophobic residue as the main structural determinant in the formation of the hydrophobic core and the canonical EGF fold. The disruption of the hydrophobic interactions leads to an alternate fold as in the case of TMEGF5.

DISCUSSION

Protein folding is a fundamental, not-yet-understood problem. It is clear from Nobel Prize Laureate C. B. Anfinsen's work in the 1960s that proteins can spontaneously refold into their native conformation. This process of refolding is mainly governed by the amino acid sequence of the protein and ensuing inherent thermodynamics. In addition, the Levinthal's paradox suggests that proteins fold into their native conformation through folding pathways. Although the specifics of folding pathways are ambiguous, it is known that folding occurs through folding intermediates, with mechanisms varying between proteins. However, all studies unanimously indicate that the amino acid sequence indeed determines protein folding.

In 2000, we and others identified a new family of conotoxins (λ/χ-conotoxins) with unique disulfide pattern and protein folding.[15−17] It has four cysteine residues in similar positions as α-conotoxins, but the disulfide linkages were C_1-C_4, C_2-C_3 in contrast to C_1-C_3, C_2-C_4 linkages of α-conotoxins. We identified two structural features, C-terminal amidation and a conserved Pro residue in intercysteine loop 1 that are exclusively found in α-conotoxins.[18,19] In contrast, λ/χ-conotoxins have a free carboxylate group and Lys or Ser residue replacing Pro. Experimentally, we showed that, when the C-terminal amidation is removed and the conserved Pro is substituted by Lys, α-conotoxin ImI preferentially folds into the C_1-C_4, C_2-C_3 isoform. Similarly, when the C-terminal is amidated and Lys is substituted by Pro in λ/χ-conotoxin CMrVIA, it preferentially folds into the C_1-C_3, C_2-C_4 isoform. Thus, we identified key determinants for alternate folds in conotoxins by comparing the structures of two classes, which differ in their disulfide pairing.

In this article, we have chosen TMEGF4 (canonical; C_1-C_3, C_2-C_4, C_5-C_6) and TMEGF5 (noncanonical; C_1-C_2, C_3-C_4, C_5-C_6) to serve as excellent models to understand the folding of EGF domains. We speculated that the segment encompassing C_1 to C_4 may determine two distinct disulfide pairings as the third disulfide pair (C_5-C_6) is analogous. Air oxidation and redox reagent-mediated oxidation studies of truncated TMEGF4 and TMEGF5 suggested that the structural determinants of both domains lie locally within their N-terminal segments. The disruption of side-chain interactions in the oxidative folding studies performed in the presence of 6 M Gn·HCl changed the folding tendency of t-TMEGF4 from its native C_1-C_3, C_2-C_4 fold into the C_1-C_2, C_3-C_4 fold. On the contrary, the disruption of side-chain interactions did not affect the folding tendency of t-TMEGF5. These observations suggested that side-chain interactions are needed to guide the fold of EGF domains toward its canonical C_1-C_3, C_2-C_4 conformer. In the absence of side-chain interactions, the peptide adopts a default conformation with C_1-C_2, C_3-C_4 pairings. These interactions are hydrophobic in nature as t-TMEGF4 folds into the canonical C_1-C_3, C_2-C_4 isoform in higher proportion when the hydrophobic effect was increased in the presence of high salt concentration. We rule out electrostatic forces as the dominant force determining the EGF fold. In t-TMEGF4, there are several oppositely charged residues, Glu3/Asp6/Glu14 and His1/Arg10. If electrostatic interactions contributed significantly to the formation of the C_1-C_3, C_2-C_4 fold, disruption of these forces by high salt

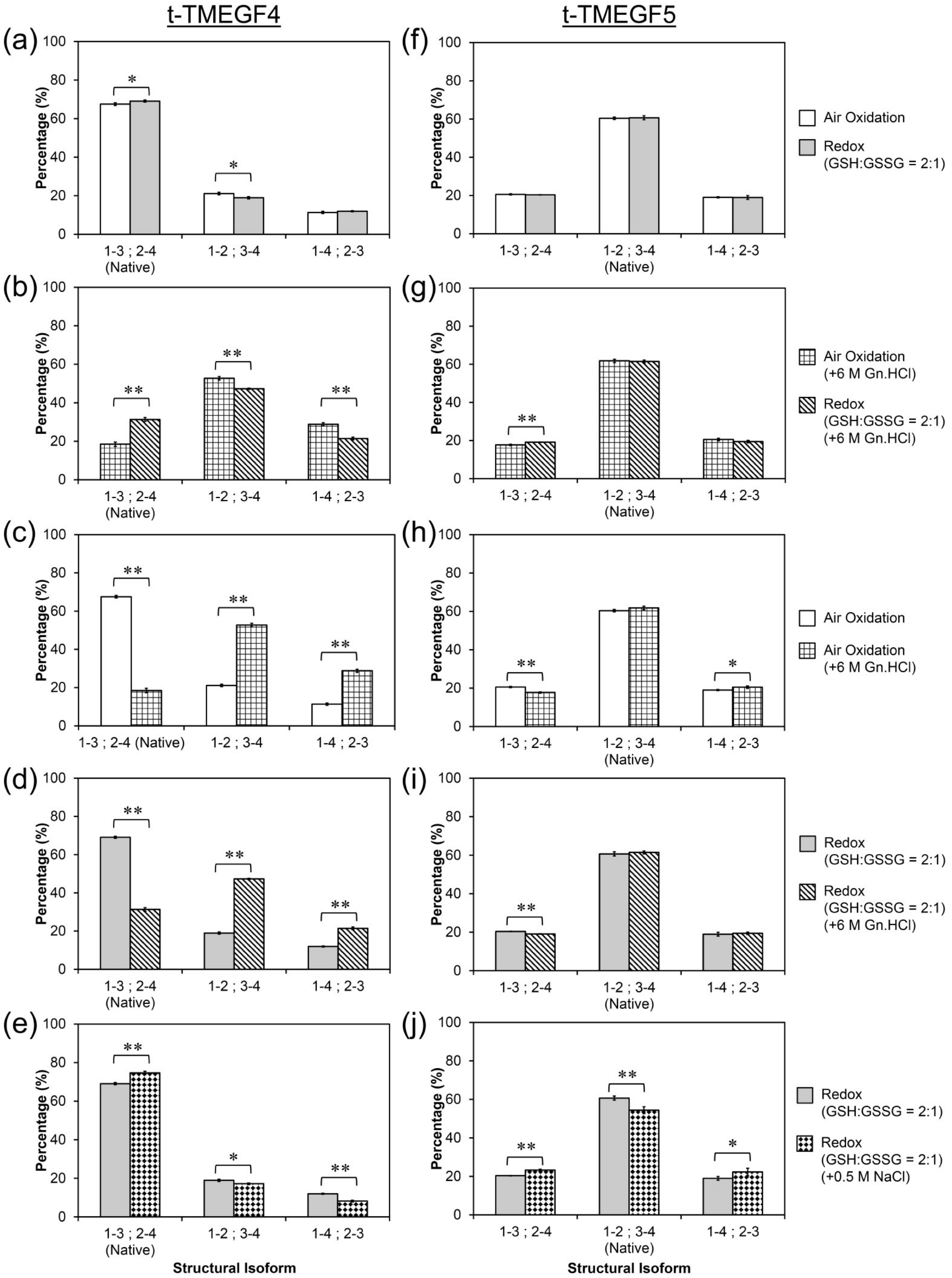


Figure 3. Folding propensities of t-TMEGF4 and t-TMEGF5 peptides. Pairwise comparisons were made to ascertain the effect of air oxidation and redox reagent-mediated oxidation, in the absence or presence of additives (6 M Gn·HCl or 0.5 M NaCl), on the proportions of structural isoforms obtained. Student's *t* test (independent samples) using arcsine transformed-values were used for the calculation of probability (p)-values. Difference in proportion between corresponding structural isoforms is deemed to be significant when the p-value is less than 0.05 (one-tailed, $n = 3$). A single asterisk (*) indicates $0.01 \leq p < 0.05$, while a double-asterisk (**) indicates $p < 0.01$: 1−3, 2−4; 1−2, 3−4; and 1−4, 2−3 indicate the disulfide connectivity of the three structural isoforms. Native isoforms are identified.

	UniProt ID	Residue	Sequence
(a)	t-EGF		
	P05979	32 - 58	-PVNPC---CYYPCQHQG---ICVRF-GLDR---**Y**QCD
	Q05769	18 - 43	--ANPC---CSNPCQNRG---ECMST-GFDQ---**Y**KCD
	P35442	549 - 575	-PVDGC---LSNPCFPGA---QCSSF-PDGS---**W**SCG
	P07204	365 - 389	-PVDPC---FRANC--EY---QCQPL-NQTS---**Y**LCV
	P08709	106 - 131	-DGDQC---ASSPCQNGG---SCKD--QLQS---**Y**ICF
	P00742	86 - 111	-DGDQC---ETSPCQNQG---KCKD--GLGE---**Y**TCT
	P01133	972 - 1002	-SDSECPLSHDGYCLHDG---VCMYIEALDK---**Y**ACN
	P01132	978 - 1008	-SYPGCPSSYDGYCLNGG---VCMHIESLDS---**Y**TCN
	Q6ULR6	44 - 68	--TASC---QDMSCSKQG---ECL--ETIGN---**Y**TCS
	P35442	590 - 620	-DLDECAL-VPDICFSTSKVPRCVN--TQPG---**F**HCL
	P46531	412 - 439	-DVDECSL-GANPCEHAG---KCIN--TLGS---**F**ECQ
	P46531	452 - 477	-DVNEC---VSNPCQNDA---TCLD--QIGE---**F**QCI
	P46531	490 - 515	-NTDEC---ASSPCLHNG---RCLD--KINE---**F**QCE
	P10493	384 - 411	-SQQTCAN-NRHQCSVHA---ECRD--YATG---**F**CCR
	Q02297	177 - 211	SHLVKCAEKEKTFCVNGG---ECFMVKDLSNPSR**Y**LCK
			* * * : *
(b)	t-TMEGF4		
	P07204	365 - 389	-PVDPC---FRANC--EY---QCQP-LNQTS---**Y**LCV
	P06579	139 - 163	-PVDPC---FDNNC--EY---QCQP-VGRSE---**H**KCI
	B3STX8	365 - 389	-PVDPC---FGTDC--EY---ECQV-VGRTG---**Y**RCV
	P15306	364 - 388	-LLDPC---FGSNC--EF---QCQP-VSPTD---**Y**RCI
	Q5W7P8	366 - 390	-PVDPC---FGSKC--EY---QCQP-VSQTD---**Y**RCI
	O35370	364 - 388	-QLDPC---FRSKC--EY---QCQP-VNSTH---**Y**NCI
	Q8HZ48	366 - 389	-PLDPC---FGTNC--EY---QCLP-LG-QN---**Y**RCI
			:*** * .* *: :* :. : *:
	t-TMEGF5		
	P07204	406 - 426	-MF--C---NQTAC--PA---DCDP-NTQ-----**A**SCE
	P06579	180 - 200	-MF--C---NQTSC--PA---DCDP-HYP-----**T**ICR
	B3STX8	406 - 426	-MF--C---NQTSC--PA---DCDP-NKQ-----**D**SCQ
	P15306	405 - 425	-MF--C---NETSC--PA---DCDP-NSP-----**T**VCE
	Q5W7P8	407 - 427	-MF--C---NQTAC--PA---DCDP-NSP-----**T**SCQ
	O35370	405 - 425	-MF--C---NETSC--PA---DCDP-NSP-----**S**FCQ
	Q8HZ48	406 - 426	-MF--C---NQTTC--PA---DCDP-NYP-----**S**TCL
			** * *:*:* ** **** : *

Figure 4. Sequence alignment of t-EGF domains. Sequence alignment of (a) truncated EGF (t-EGF) domains (encompassing C_1 to C_4) from various proteins and (b) t-TMEGF4/5 from various organisms. Information on the identity of the proteins and their associated EGF domains can be found in Supplementary Table S2.

concentration would have reduced the proportion of its native fold. It is also important to note that no salt bridges were observed between these residues. In t-TMEGF5, there are three acidic residues (Asp55, Asp57, and Glu65). If electrostatic repulsions between these residues are responsible for the altered disulfide pairings, the masking of these repulsions by high salt concentration would have reverted the folding back to the canonical fold. However, only a modest increase in the C_1–C_3, C_2–C_4 isoform was observed, with t-TMEGF5 still preferentially folding into its native C_1–C_2, C_3–C_4 fold. Taken together, these observations support the conclusion that hydrophobic interactions is crucial for the canonical EGF fold.

Since EGF domains are structurally conserved modular units with diverse functionality, the positions of these hydrophobic structural determinants have to be conserved to maintain the overall canonical fold while accommodating varied functional residues. To identify these structural determinants, we aligned EGF domains from various proteins. Despite the phylogenetic distances and functional differences, all the EGF domains contain a conserved hydrophobic residue in the penultimate position in intercysteine loop 3 (Figure 4, Supplementary Figure S2). This hydrophobic residue is not present in mammalian TMEGF5, which has distinct disulfide pairing and protein fold. The three-dimensional structures of canonical EGF domains show that this conserved hydrophobic residue is in hydrophobic contacts with residues in intercysteine loop 1 (Supplementary Figure S3). Although it is not clear whether these interactions occur in transition states of folding intermediates, these contacts are needed for the C_1–C_3, C_2–C_4 fold as these interactions bring C_4 and C_2 in close proximity. Therefore, we speculated that any disruption of these contacts would destabilize the hydrophobic core and hence the C_1–C_3, C_2–C_4 structure, to create the more loosely packed C_1–C_2, C_3–C_4 structure.

In TMEGF4, the hydrophobic interactions occur between Ala11 and Tyr25. However, this contact is not present in the equivalent positions in TMEGF5. To experimentally test the hydrophobic core hypothesis, we disrupted the interaction by substituting Tyr25 with a hydrophilic Thr residue. t-TMEGF4-Y25T preferentially folded into the noncanonical C_1–C_2, C_3–C_4 isoform, and this was accompanied by a sharp drop in the canonical C_1–C_3, C_2–C_4 isoform. Thus, these results strongly suggest that the hydrophobic residue in the penultimate position in intercysteine loop 3 is the key structural determinant that determines the C_1–C_3, C_2–C_4 fold. This conserved residue nucleates the hydrophobic core and acts as the lynch pin.

It is relatively easy to disrupt this hydrophobic core in t-TMEGF4 by introducing a Y25T substitution. However, it is probably harder to create a hydrophobic core in t-TMEGF5 by mere introduction of a hydrophobic residue at its equivalent position as it requires hydrophobic-interacting partners in intercysteine loop 1. TMEGF5 might have acquired several other substitutions through evolution to commit it to the noncanonical C_1–C_2, C_3–C_4, and C_5–C_6 fold. This is further supported by the observations that the relative proportion of

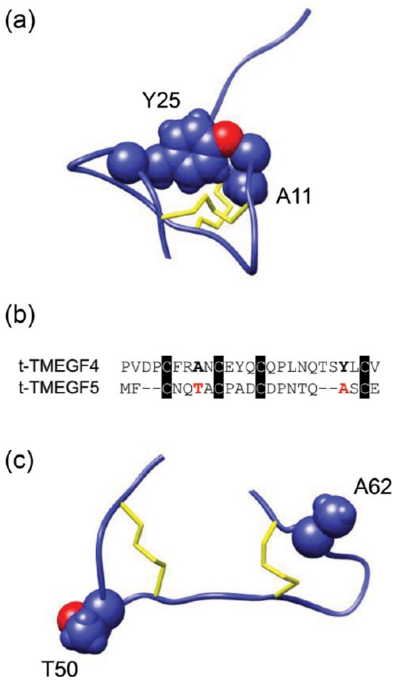

(a)
Y25
A11
(b)
t-TMEGF4 PVDPCFRANCEYQCQPLNQTSYLCV
t-TMEGF5 MF--CNQTACPADCDPNTQ--ASCE
(c)
A62
T50

Figure 5. Role of the conserved hydrophobic residue in EGF domain folding. (a) Model of t-TMEGF4 showing the conserved hydrophobic residue of Y25 interacting with A11 of the first intercysteine loop. (b) Alignment of t-TMEGF4 and t-TMEGF5 sequences. Cys residues are highlighted. Residues in t-TMEGF4 involved in the formation of hydrophobic nucleus are shown in bold. These residues are replaced by less hydrophobic residues in t-TMEGF5, shown in red. (c) Model of t-TMEGF5 showing lack of interaction between T50 and A62. The models of these segments were extracted from PDB: 1DQB. Positions of residues are labeled in accordance to the position numbers used in the PDB file.

the noncanonical C_1−C_2, C_3−C_4 conformer in t-TMEGF4 when folded in the presence of 6 M Gn·HCl or in t-TMEGF4 (Y25T) when folded under normal oxidative conditions did not reach as high as that of t-TMEGF5. Therefore, t-TMEGF5 may possess additional specific structural determinants for the noncanonical C_1−C_2, C_3−C_4 fold.

■ CONCLUSIONS

The EGF domains are commonly found in various classes of proteins and play a crucial role in diverse functions. By systematic studies of two closely related thrombomodulin domains with canonical and noncanonical disulfide pairings, we identified a conserved hydrophobic residue as the key structural determinant that plays a crucial role in determining the domain fold. We have shown that the hydrophobic core in EGF domains are mediated through the interaction between this highly conserved hydrophobic residue in intercysteine loop 3 and some hydrophobic residues in intercysteine loop 1, with the disruption of this hydrophobic core leading to domains with alternate disulfide pairings and domain fold. Protein isoforms that differ in disulfide pairings help in the identification of specific structural determinants that play a crucial role in determining the protein fold.

■ METHODS

Materials. Standard 9-fluorenylmethoxycarbonyl (Fmoc)-L-amino acid hydroxyl derivatives were purchased from AnaSpec, Inc. Novasyn TGR resin, *N*,*N*-dimethylformamide (DMF), trifluoroacetic acid (TFA), piperidine, and *N*,*N*-diisopropylethylamine (DIPEA) were purchased from Merck KGaA. *N*-Methyl-2-pyrrolidone (NMP) was purchased from Tokyo Chemical Industry. 1,2-Ethanedithiol (EDT) and triisopropylsilane (TIS) were obtained from Sigma Aldrich Co. LLC. Jupiter Proteo 4 *μ* 90 Å (15 × 250 mm) column and Kinetex PFP 2.6 *μ* 100 Å (4.6 × 100 mm) column were purchased from Phenomenex, Inc. Cosmosil Cholester 5 *μ* 120 Å (4.6 × 250 mm) column was purchased from Nacalai Tesque, Inc. All other chemicals and reagents used were of analytical grade.

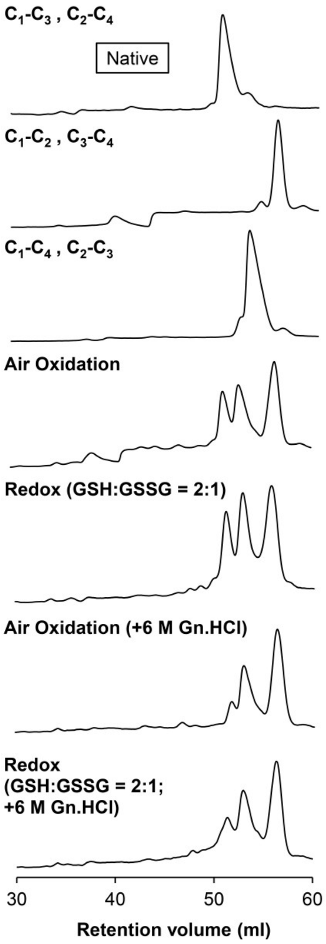
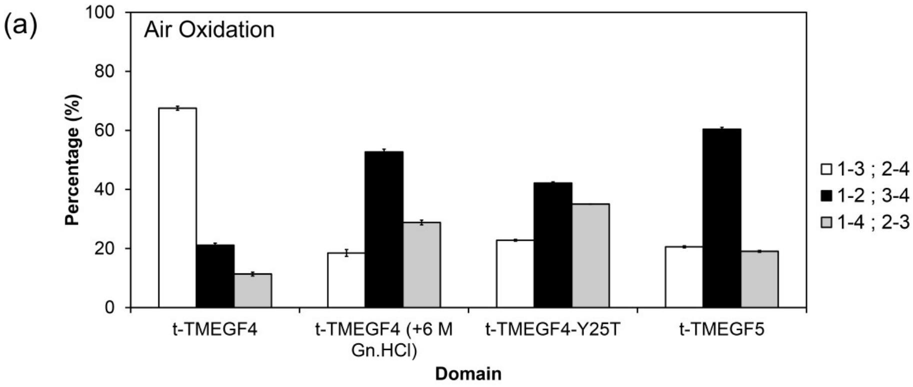


Figure 6. Oxidative folding of t-TMEGF4-Y25T peptide. Structural isoforms obtained were identified and quantified by measuring their corresponding peak areas. The folding conditions used to study t-TMEGF4-Y25T were air oxidation and redox reagent-mediated oxidation, in the absence or presence of 6 M Gn·HCl.

Peptide Synthesis. All peptides were synthesized using manual Fmoc-solid phase peptide synthesis on the Novasyn TGR resin. The coupling step was performed in DMF:NMP (2:1) with 5 times excess of amino acid derivatives activated *in situ* by 4.9 times excess of HATU and 10 times excess of DIPEA. The removal of Fmoc-moiety was achieved using a solution of 20% (v/v) piperidine in DMF.

The assembled peptides were cleaved from the resin using a cocktail of TFA/EDT/TIS/water (94:2.5:1:2.5% v/v) and precipitated using ice-cold diethyl-ether. The crude peptides were then purified using a Jupiter Proteo 4 *μ* 90 Å (15 × 250 mm) column on a ÄKTA purifier system (GE Healthcare). Fractions containing the target peptide were identified using electrospray ionization mass spectrometry (ESI-MS) on an API-300 LC/MS/MS system (Perkin-Elmer, Inc.).

Regioselective Synthesis of Structural Isoforms. By placing *S*-trityl (Trt) or *S*-acetamidomethyl (Acm)-protected cysteine residues at specific positions along the peptide chain, orthogonal protection of cysteine side-chains were used to generate structural isoforms (Supplementary Figure S1). Cysteine residues involved in the first



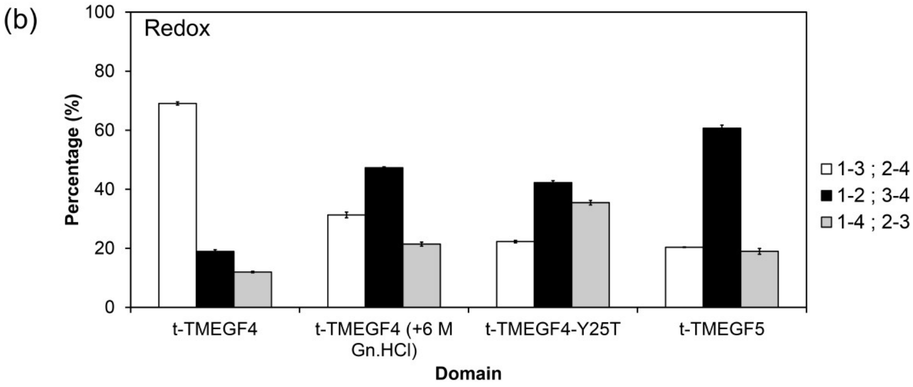


Figure 7. Comparison of folding propensities of t-TMEGF4-Y25T with relevant peptides. The dominant isoform of t-TMEGF4 obtained under normal oxidative conditions is the canonical EGF-domain (C_1−C_3, C_2−C_4) fold, while the dominant isoform obtained for t-TMEGF4 in the presence of 6 M Gn·HCl, t-TMEGF4-Y25T, and t-TMEGF5 is the noncanonical C_1−C_2, C_3−C_4 fold: 1−3, 2−4; 1−2, 3−4; and 1−4, 2−3 indicate the disulfide connectivity of the three structural isoforms.

disulfide bridge were protected with the acid labile Trt-group, which were removed upon TFA treatment in the peptide synthesis cleavage step. The disulfide bridge between the two free cysteine residues was then formed by stirring the Cys(Acm)-containing peptides (at a concentration of 0.3 mM) in a 0.1 M Tris-HCl, pH 7.5 buffer containing 10% (v/v) acetonitrile (ACN) and 20% (v/v) DMSO for 24 h. Using ESI-MS, complete formation of the first disulfide bridge was confirmed by the reduction of two mass units. The peptides were purified on a Jupiter Proteo 4 μ 90 Å (15 × 250 mm) column and lyophilized before proceeding to the next oxidation step.

For the formation of the second disulfide bridge, the remaining two Acm-protected cysteine residues were simultaneously deprotected and oxidized by adding solid iodine (5 eq/Acm) and HCl (1.5 eq/Acm) to a solution containing 0.6 mM peptide in 80% (v/v) acetic acid. The solution was stirred vigorously for 1 h before quenching with 1 M ascorbic acid dropwise until the solution become colorless. Following purification, ESI-MS was used to identify fractions containing the completely oxidized peptides (marked by a mass reduction of 144 Da).

Oxidative Folding Studies. Oxidative folding studies were conducted using fully deprotected (i.e., all cysteine residues derived from Cys(Trt) derivative) and reduced peptides. For air oxidation, 0.1 mM of peptide was dissolved in 0.1 mM Tris-HCl, pH 8.0, containing 10% (v/v) ACN. The solution was stirred in an open atmosphere, with the progress of the reaction monitored using the Ellman's test. When the reaction was deemed complete, the pH of the solution was adjusted to pH 2 using concentrated HCl. For air oxidation in the presence of denaturant, 6 M Gn·HCl was included in the buffer.

For oxidation using glutathione redox system, 0.1 mM of peptide was dissolved in 0.1 M Tris-HCl, pH 8.0, containing 1 mM EDTA, 2 mM reduced glutathione, 1 mM oxidized glutathione, and 10% (v/v) ACN. The solution was then purged with nitrogen gas before the reaction tube was sealed. The reaction was allowed to proceed with stirring for 48 h before the pH of the solution was adjusted to pH 2. For redox reagent-mediated oxidation in the presence of denaturant or high salt content, 6 M Gn·HCl or 0.5 M NaCl was included in the buffer, respectively.

Structural isoforms of t-TMEGF4 and t-TMEGF4-Y25T obtained from the oxidative folding reactions were separated using the Cosmosil Cholester 5 μ 120 Å (4.6 × 250 mm) column. For t-TMEGF4, a segmented gradient elution method involving TFA as the counterion (constant concentration of 0.1% v/v) and methanol (MeOH) as the organic modifier (maximum 60% v/v) was used. For t-TMEGF4-Y25T, a segmented gradient elution method involving heptafluorobutyric acid (HFBA) as the counterion (constant concentration of 10 mM) and MeOH as the organic modifier (maximum 80% v/v) was used.

Structural isoforms of t-TMEGF5 obtained were separated using the Kinetex PFP 2.6 μ 100 Å (4.6 × 100 mm) column. A segmented gradient elution method involving HFBA as the counterion (constant concentration of 10 mM) and MeOH as the organic modifier (maximum 80% v/v) was used.

The amount of each structural isoform obtained is quantified by measuring the peak area of its corresponding peak in the chromatogram. The peak area was calculated using the peak integration function of the UNICORN protein purification software (GE Healthcare). Skim procedures were applied when deemed necessary to improve the accuracy of calculations.

Statistical Analysis. All oxidative folding experiments were performed in triplicates. The amount of each structural isoform obtained was expressed as percentage values before the average and standard deviation values were calculated.

The Student's *t* test (for independent samples) was used to test for significant differences in the proportion of structural isoforms obtained from two different oxidative folding conditions. It should be noted that, for a parametric test, the direct input of percentage data is not recommended. Thus, in accordance to a solution recommended by Zar,[20] an arcsine transformation was performed on all percentage values (from each replicate) before the statistical test was performed.

■ ASSOCIATED CONTENT

Supporting Information

Regioselective synthesis of t-TMEGF4 and t-TMEGF5; sequence alignment of t-EGF domains; residues interacting with the conserved hydrophobic residue in canonical EGF domains; observed versus theoretical mass of t-TMEGF4 and t-TMEGF5 structural isoforms; identity of proteins and their associated EGF domains. This material is available free of charge *via* the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: dbskinim@nus.edu.sg.

Present Address

‖School of Life Sciences and Chemical Technology, Ngee Ann Polytechnic, Singapore 599489, Singapore.

Author Contributions

A.S.A.N. performed all of the studies, and R.M.K. contributed to the concept and experimental planning. Both authors contributed to writing the manuscript and creating the figures.

Notes

The authors declare no competing financial interest.

■ REFERENCES

(1) Anfinsen, C. B. (1972) The formation and stabilization of protein structure. *Biochem. J.* *128*, 737−749.

(2) Anfinsen, C. B., Haber, E., Sela, M., and White, F. H., Jr. (1961) The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proc. Natl. Acad. Sci. U.S.A.* *47*, 1309−1314.

(3) White, C. E., Hunter, M. J., Meininger, D. P., Garrod, S., and Komives, E. A. (1996) The fifth epidermal growth factor-like domain of thrombomodulin does not have an epidermal growth factor-like disulfide bonding pattern. *Proc. Natl. Acad. Sci. U.S.A.* *93*, 10177−10182.

(4) Esmon, C. T., Esmon, N. L., and Harris, K. W. (1982) Complex formation between thrombin and thrombomodulin inhibits both thrombin-catalyzed fibrin formation and factor V activation. *J. Biol. Chem.* *257*, 7944−7947.

(5) Esmon, C. T., and Owen, W. G. (1981) Identification of an endothelial cell cofactor for thrombin-catalyzed activation of protein C. *Proc. Natl. Acad. Sci. U.S.A.* *78*, 2249−2252.

(6) White, C. E., Hunter, M. J., Meininger, D. P., White, L. R., and Komives, E. A. (1995) Large-scale expression, purification and characterization of small fragments of thrombomodulin: the roles of the sixth domain and of methionine 388. *Protein Eng.* *8*, 1177−1187.

(7) Meininger, D. P., Hunter, M. J., and Komives, E. A. (1995) Synthesis, activity, and preliminary structure of the fourth EGF-like domain of thrombomodulin. *Protein Sci.* *4*, 1683−1695.

(8) Wood, M. J., Sampoli Benitez, B. A., and Komives, E. A. (2000) Solution structure of the smallest cofactor-active fragment of thrombomodulin. *Nat. Struct. Biol.* *7*, 200−204.

(9) Sampoli Benitez, B. A., Hunter, M. J., Meininger, D. P., and Komives, E. A. (1997) Structure of the fifth EGF-like domain of thrombomodulin: An EGF-like domain with a novel disulfide-bonding pattern. *J. Mol. Biol.* *273*, 913−926.

(10) Hunter, M. J., and Komives, E. A. (1995) Thrombin-binding affinities of different disulfide-bonded isomers of the fifth EGF-like domain of thrombomodulin. *Protein Sci.* *4*, 2129−2137.

(11) Saez, G., Thornalley, P. J., Hill, H. A., Hems, R., and Bannister, J. V. (1982) The production of free radicals during the autoxidation of cysteine and their effect on isolated rat hepatocytes. *Biochim. Biophys. Acta* *719*, 24−31.

(12) Volkin, D. B., Mach, H., Middaugh, C. R. (1995) Degradative Covalent Reactions Important to Protein Stability, in *Protein Stability and Folding: Theory and Practice* (Shirley, B. A., Ed.) pp 35−64, Humana Press, Totowa, NJ.

(13) Shields, P. A., and Farrah, S. R. (1983) Influence of salts on electrostatic interactions between poliovirus and membrane filters. *Appl. Environ. Microbiol.* *45*, 526−531.

(14) Melander, W. R., Corradini, D., and Horvath, C. (1984) Salt-mediated retention of proteins in hydrophobic-interaction chromatography. Application of solvophobic theory. *J. Chromatogr.* *317*, 67−85.

(15) Balaji, R. A., Ohtake, A., Sato, K., Gopalakrishnakone, P., Kini, R. M., Seow, K. T., and Bay, B. H. (2000) Lambda-conotoxins, a new family of conotoxins with unique disulfide pattern and protein folding. Isolation and characterization from the venom of *Conus marmoreus*. *J. Biol. Chem.* *275*, 39516−39522.

(16) McIntosh, J. M., Corpuz, G. O., Layer, R. T., Garrett, J. E., Wagstaff, J. D., Bulaj, G., Vyazovkina, A., Yoshikami, D., Cruz, L. J., and Olivera, B. M. (2000) Isolation and characterization of a novel conus peptide with apparent antinociceptive activity. *J. Biol. Chem.* *275*, 32391−32397.

(17) Sharpe, I. A., Gehrmann, J., Loughnan, M. L., Thomas, L., Adams, D. A., Atkins, A., Palant, E., Craik, D. J., Adams, D. J., Alewood, P. F., and Lewis, R. J. (2001) Two new classes of conopeptides inhibit the alpha1-adrenoceptor and noradrenaline transporter. *Nat. Neurosci.* *4*, 902−907.

(18) Kang, T. S., Radic, Z., Talley, T. T., Jois, S. D., Taylor, P., and Kini, R. M. (2007) Protein folding determinants: structural features determining alternative disulfide pairing in alpha- and chi/lambda-conotoxins. *Biochemistry* *46*, 3338−3355.

(19) Kang, T. S., Vivekanandan, S., Jois, S. D., and Kini, R. M. (2005) Effect of C-terminal amidation on folding and disulfide-pairing of alpha-conotoxin ImI. *Angew. Chem., Int. Ed.* *44*, 6333−6337.

(20) Zar, H. J. (1984) *Biostatistical Analysis*, Prentice Hall International, Upper Saddle River, NJ.