Eur. J. Biochem. **271**, 4229–4240 (2004) © FEBS 2004 doi:10.1111/j.1432-1033.2004.04363.x

Folding of epidermal growth factor-like repeats from human tenascin studied through a sequence frame-shift approach

Francesco Zanuttin, Corrado Guarnaccia, Alessandro Pintar and Sándor Pongor

International Centre for Genetic Engineering and Biotechnology (ICGEB), Protein Structure and Bioinformatics Group, Trieste, Italy

In order to investigate the factors that determine the correct folding of epidermal growth factor-like (EGF) repeats within a multidomain protein, we prepared a series of six peptides that, taken together, span the sequence of two EGF repeats of human tenascin, a large protein from the extracellular matrix. The peptides were selected by sliding a window of the average length of tenascin EGF repeats over the sequence of EGF repeats 13 and 14. We thus obtained six peptides, EGF-f1 to EGF-f6, that are 33 residues long, contain six cysteines each, and bear a partial overlap in the sequence. While EGF-f1 corresponds to the native EGF-14 repeat, the others are frame-shifted EGF repeats. We carried out the oxidative folding of these peptides *in vitro*, analyzed the reaction mixtures by acid trapping followed by LC-MS, and isolated some of the resulting products. The oxidative folding of the native EGF-14 peptide is fast, produces a single three-disulfide species with an EGF-like disulfide topology and a marked difference in the RP-HPLC retention time compared with the starting product. On the contrary, frame-shifted peptides fold more slowly and give mixtures of three-disulfide species displaying RP-HPLC retention times that are closer to those of the reduced peptides. In contrast to the native EGF-14, the three-disulfide products that could be isolated are mainly unstructured, as determined by CD and NMR spectroscopy. We conclude that both kinetics and thermodynamics drive the correct pairing of cysteines, and speculate about how cysteine mispairing could trigger disulfide reshuffling *in vivo*.

Keywords: EGF; folding; disulfide; extracellular proteins.

Tenascin-C [1–3] is a large extracellular matrix glycoprotein expressed during embryonic development and in proliferative processes such as wound healing and tumorigenesis. Although the function of tenascin is not fully understood, careful studies on tenascin-C-deficient mice recently highlighted the function of tenascin in hematopoiesis [4] and identified behavioral abnormalities that point to a role of tenascin in the development and maintenance of proper brain chemistry [5]. The cloning of tenascin unraveled its modular architecture [6–8]. The N-terminal region, which is responsible for tenascin oligomerization, is followed by a series of 14 epidermal growth factor-like (EGF) tandem repeats, 15 fibronectin type III domains and a C-terminal fibrinogen-like domain. Several studies attempted to map the different biological activities of tenascin to selected domains. Recently, a role of tenascin EGF repeats as immobilized, low affinity ligands for the EGF receptor (EGFR) has been proposed [9], stemming from the observation that selected EGF repeats of tenascin-C bind and directly activate EGFR and induce mitogenesis in mouse fibroblasts.

Correspondence to A. Pintar, International Centre for Genetic Engineering and Biotechnology (ICGEB), Protein Structure and Bioinformatics Group, AREA Science Park, Padriciano 99, I-34012 Trieste, Italy. Fax: +39 040 226555, Tel.: +39 040 3757354, E-mail: pintar@icgeb.org

Abbreviations: EGF, epidermal growth factor; TBTU, *O*-(benzotriazol-1-yl)-1,1,3,3-tetramethyluronium tetrafluoroborate; Trt, trityl.

(Received 28 June 2004, revised 26 August 2004, accepted 8 September 2004)

EGF domains [10] are 30–50 residue long repeats characterized by the strict conservation of six cysteine residues that form three disulfide bonds with the topology 1–3, 2–4, 5–6. The common structural feature of EGF domains is a two-stranded β-sheet from which the three disulfide bonds depart to connect the N- and C-terminal loops, to make a rather compact structure [11]. Beside the six cysteines, a wide variability in the length and composition of the stretches connecting the cysteines has been observed. Probably because of its capability to accommodate very different sequences on a common scaffold, the EGF domain is one of the most frequently employed building blocks in modular proteins. EGF domains are found in more than 300 human extracellular proteins [12]. As EGF domains occur very frequently as multiple tandem repeats, the total number in human proteins exceeds 4000 [12]. The oxidative folding *in vitro* of human EGF [13], as well as the folding of other small three-disulfide domains [14], has been studied in detail. The only general conclusion is that while the disulfide topology of the final product is well conserved, the oxidative folding pathway of EGF domains is rather complex and unpredictable [14]. In human EGF, the rapid formation of the second and third disulfide bonds leads to an intermediate species that accumulates and acts as a kinetic trap [13]; the native product, however, is not developing through the formation of the third disulfide bond, which is slow, but rather from the scrambling of disulfide bonds in other three-disulfide, non-native products. The disulfide bonds lock the conformation of the protein into a stable structure [15] even though not all the disulfide bridges are equally important for the maintenance of the 3D structure [16]. The human EGF precursor protein

itself is in fact expressed as a 150–180 kDa multidomain type I membrane protein [17] containing nine EGF-like domains. The soluble 53 residue EGF corresponds to the ninth EGF domain in the precursor, from which it is released by proteolytic cleavage.

A long stretch of 14 tandem EGF repeats is found in the N-terminal region of human tenascin [8]. The 14 EGF repeats, which are encoded by a single exon [7], show a variable degree of similarity within each other, which ranges from 35 to 74% identity, and have the peculiarity to be only 31 residues long, with a spacing of 25 residues between the first and the last cysteine. The correct pairing of cysteines to form disulfide bridges is critical to reach the final native fold, and we wondered, on the one hand, what determines the pairing of cysteines to give the correct disulfide bond pattern within each repeat and, on the other hand, what drives this topology to be repeated in the same frame along the amino acid sequence of a multirepeat protein. In fact, little is known about the folding of large modular proteins that are targeted to the extracellular environment, and the inherent complexity of oxidative folding in cysteine-rich proteins requires a simple model system that can be studied in detail by physico-chemical methods.

With this purpose, we prepared, by solid-phase synthesis, a series of six peptides (Fig. 1) that, taken together, span the sequence of the two last EGF repeats of human tenascin, EGF-13 and EGF-14. The peptides were designed by selecting a window of 33 amino acids, which corresponds to the average length of the tenascin EGF repeats, and sliding this window over the amino acid sequence of tenascin EGF repeats 13 and 14 (residues 560–622). The window was slid by one cysteine at each step, thus obtaining six peptides named EGF-f1 to EGF-f6, that are all 33 residues long, contain six cysteines, and bear a partial overlap in their sequences. While EGF-f1 corresponds to the putative EGF-14 repeat, the others are frame-shifted EGF repeats. We carried out the oxidative folding of these peptides in the presence of a redox couple, analyzed the reaction mixture by acid trapping followed by LC-MS, compared the different folding profiles, and characterized some of the three-disulfide products that are formed.

We discuss the significance of the frame-shift approach in terms of the kinetic and thermodynamic aspects that drive the correct folding of EGF repeats within multidomain proteins, and in relation to the folding *in vivo* of disulfide-rich proteins.

Experimental procedures

Reagents

Fmoc-protected amino acids were purchased from Chem-Impex International (Wood Dale, IL, USA), Fluka (Buchs, Switzerland), Advanced Biotech Italia (Seveso, Italy) and NovaBiochem (Darmstadt, Germany). TentaGel S trityl (Trt) resins loaded with the required Fmoc-protected amino acids (Fluka) were chosen as solid supports. The resin capacity ranged from 0.18 to 0.2 mmol·g^{-1}. Synthesis-grade reagents employed in the peptide synthesis were from Biosolve LTD (Valkenswaard, the Netherlands) except 2,6-dimethylpyridine and diisopropylethylamine, which were obtained from Aldrich (Steinheim, Germany).

Chemicals used in cleavage and deprotection steps were from Aldrich and Fluka, trifluoroacetic acid from Biosolve. HPLC grade acetonitrile for chromatographic separations was obtained from Riedel-deHaen (Seelze, Germany). Endopeptidase AspN (27750 U·mg^{-1}) and thermolysin (8560 U·mg^{-1}) were from Calbiochem (Darmstadt, Germany).

Peptide synthesis

The 33 amino acid peptide corresponding to residues 590–622 of human tenascin-C (Swiss-Prot: TENA_HUMAN), EGF-14 (Fig. 1), was synthesized by solid-phase Fmoc based strategy. The synthesis was automatically performed

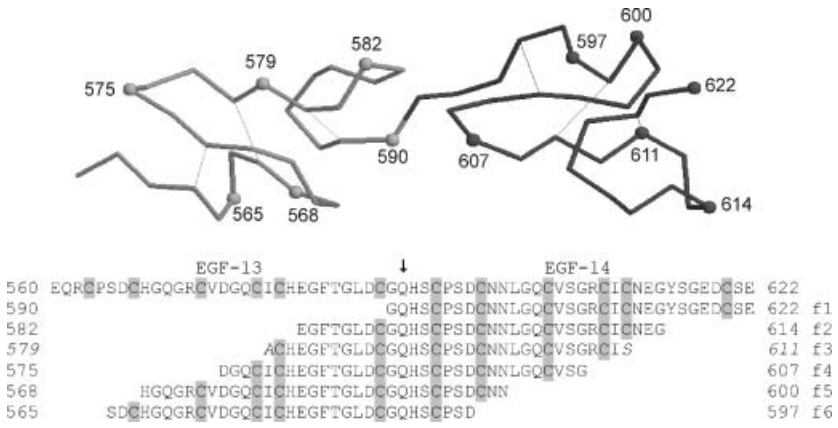

Fig. 1. Amino acid sequence of human tenascin EGF-repeats 13 and 14. Amino acid sequence of human tenascin (Swiss-Prot: TENA_HUMAN, residues 560–622) EGF-repeats 13 and 14, and of the synthesized peptides; f1 corresponds to EGF-14, while f2–f6 correspond to the different frame-shifted peptides. Cysteines are highlighted in gray, non-native residues are in italics, the limit between the two EGF-repeats is shown by an arrow. A model of the tandem repeats is also shown on top as a Cα trace. After a search for a suitable template with 3D-PSSM [38] the model was built by MODELLER [39] using the structure of an EGF pair from fibrillin (PDB: 1emn) as template. Because of the low sequence similarity between tenascin and fibrillin EGF repeats (38% identity) the model is only approximate. To map the synthesized peptides over the structure, peptide limits are pinpointed by spheres and labeled by residue number.

with a PS3 Protein Technology (Tucson, AZ, USA) synthesizer on a 0.07-mmol scale. The Fmoc protected amino acid, the coupling reagent [TBTU; *O*-(benzotriazol-1-yl)-1,1,3,3-tetramethyluronium tetrafluoroborate], and the base (diisopropylethylamine), in a ratio of 1 : 1 : 2, were dissolved in dimethylformamide using a four molar excess in respect to the initial resin substitution, to give a final amino acid concentration of 0.3 M. Each coupling step took 45 min from S2 to C16 and 1.5 h from C16 to G33. Fmoc deprotection was carried out in 20% (v/v) piperidine in dimethylformamide for 5 min and the reaction repeated twice. Cysteine residues were added manually as *N*-α-Fmoc-*S*-trityl-L-cysteine pentafluorophenyl ester [Fmoc-Cys(Trt)-OPfp] dissolved in dimethylformamide in a 2-h reaction in order to avoid cysteine racemization [18]. The side chain-protected peptide-resin was washed with dichloromethane, dried, cleaved and deprotected in 90% (v/v) trifluoroacetic acid, 5% (v/v) 1,2-ethanedithiol, 2.5% (v/v) triisopropylsilane, 2.5% (v/v) water and phenol (0.5 M) for 2 h at room temperature. The solution was filtered in order to remove the resin and trifluoroacetic acid was evaporated in vacuum. The deprotected peptide was dissolved in water, the solution extracted five times with 6–8 volumes of diethyl ether to remove scavengers, and finally freeze-dried. The crude EGF-14 peptide was purified by RP-HPLC on a Gilson chromatographic apparatus using a Zorbax 300SB-C18 9.4 × 250 mm column (Agilent) with a linear gradient of triethylammonium acetate buffer (25 mM, pH 7) and triethylammonium acetate buffer (25 mM, pH 7) in water/acetonitrile 1 : 9 (v/v).

Frame-shifted peptides, f2 to f6 (Fig. 1) were manually synthesized by standard stepwise solid-phase procedure on a 0.1-mmol scale. In f3, an Ala residue was inserted instead of Ile at the N-terminus and an extra Ser residue was added at the C-terminus to avoid the presence of bulky aliphatic residues or Cys, respectively, at the peptide ends. The coupling reactions were performed with 4 eq of the Fmoc-protected amino acids and activator (TBTU), and 8 eq of diisopropylethylamine in dimethylformamide for 1 h. Kaiser's ninhydrin test [19] was systematically applied after each coupling in order to check reaction completion. Fmoc protecting groups were removed by a 20% (v/v) solution of piperidine in dimethylformamide containing 0.1 M 1-hydroxybenzotriazole. When necessary, a second coupling reaction was made using 1*H*-benzotriazol-1-yl-oxy-tris(pyrrolidino)phosphonium hexafluorophosphate (PyBop) or *O*-(7-azabenzotriazol-1-lyl)-1,1,3,3-tetramethyluronium hexafluorophosphate as coupling reagents in dimethylformamide. The double coupling procedure was systematically adopted with cysteine amino acids. Also in this case, Fmoc-Cys(Trt)-OPfp was employed in the first reaction to minimize Cys racemization [18]. In the second coupling, a four molar excess solution of Fmoc-Cys(Trt)-OH, TBTU, 2,6-dimethylpyridine, in 1 : 1 : 2 ratio in dichloromethane/dimethylformamide 1 : 1 (v/v) was used. Peptides were cleaved from the resin and deprotected as described for EGF-14. Preparative RP-HPLC of frame-shifted peptides f2–f6 was carried out on the same Gilson chromatographic apparatus using a PrePak Cartridge 25 × 100 mm (Agilent) casted on a PrepLC Universal Base apparatus (Waters) and a Zorbax 300SB-C18 9.4 × 250 mm (Agilent). Samples were eluted using a linear gradient of water/trifluoroacetic acid 0.1% (v/v) (buffer A) and acetonitrile/trifluoroacetic acid 0.1% (v/v) (buffer B).

Analysis of all peptides was carried out on the same chromatographic system. Sample elution was followed by UV detection at 214 nm. Two Zorbax 300SB-C18 columns (Agilent) of different diameters were used: 1.0 × 150 mm, 3.5 μm and 4.6 × 150 mm, 3.5 μm with the same buffers. The identity of the peptides was checked by LC-MS (see below).

Oxidative folding

After purification by RP-HPLC in acidic conditions, the reduced and lyophilized peptides (EGF-14 and f2–f6) were dissolved in an acidic water solution [trifluoroacetic acid 0.01% (v/v)] and immediately diluted 10× in the refolding buffer [0.1 M ammonium acetate, 2 mM EDTA, Cys/cystine 20 : 1 (w/w), pH 8.5] previously flushed with nitrogen. The molar ratio between peptide cysteines and the cystine in the redox couple was 10 : 1. Comparable amounts of each peptide, as estimated by UV absorbance, were dissolved in a final volume of 5 mL and used in time course refolding experiments. Aliquots of the reaction mixtures were quenched at selected times (2.5, 5, 10, 15, 20, 30, 40, 60, 90, 120 min, 4 and 24 h) by acidification with trifluoroacetic acid to yield a final 2% (v/v) trifluoroacetic acid concentration and stored at − 80 °C. The different species were identified by LC-MS (Mass spectrometry section) analysis and quantified by peak integration of the RP-HPLC profile (UV detection at 214 nm).

The final products from the folding reactions of EGF-14, f5, and f6 were purified by RP-HPLC using a Zorbax 300SB-C18 (4.6 × 150 mm, 3.5 μm) column with the same buffers A and B.

Disulfide bond determination of EGF-14

To define the intramolecular disulfide topology, EGF-14 was treated with thermolysin and AspN. In the first reaction 40 μg of the purified peptide dissolved in a 10-mM, pH 6.0 buffer was digested with 12 μg of thermolysin for 12 h at 37 °C. A part of the digest was further incubated with AspN for 6 h at 37 °C. Peptide fragments obtained from the digestions were fractionated by RP-HPLC with a water/acetonitrile 0.1% (v/v) trifluoroacetic acid linear gradient and analyzed by LC-MS. Reactions in the absence of the enzyme and in the absence of the substrate were used as negative controls.

Mass spectrometry

MS analysis was carried out with an API 150 EX single quadrupole mass spectrometer (PE/Sciex, Thomhill, Canada) equipped with an ion spray source. The identity of the synthesized peptides was checked and the digestion mixtures analyzed by LC-MS using a Zorbax 300SB C18 column (1.0 × 150 mm, 3.5 μm) (Agilent) with a linear gradient of buffer A and B at a flow rate of 50 $\mu L \cdot min^{-1}$. The analysis was achieved in positive-ion mode. Time-course refolding experiments were followed by LC-MS in the same conditions. In order to detect all possible disulfide species, the MS

spectrum was acquired over two 20 atomic mass units wide windows centered on the *m*/*z*-values corresponding to the double and triple charged reduced peptide. The mass spectrometer was run in total ion count mode with a step of 0.1 atomic mass units and a 1.5-ms dwell time, the orifice voltage being set at 30 V. The reconstruction of the original molecular mass of the peptides was achieved using the BioMultiview software (Applied Biosystem).

NMR spectroscopy

Samples for NMR spectroscopy were prepared dissolving the lyophilized peptides in H_2O/D_2O (90/10, v/v) and adjusting the pH to ≈ 5.5 with NaOH 0.1 M. Sample concentration was ≈ 2 mM for EGF-14, ≈ 50 μM for f5b and f6b, 10–15 μM for f5a and f6a. Spectra were recorded on a Bruker Avance DRX 500 operating at a ^1H frequency of 500.12 MHz and equipped with a triple resonance, *z*-axis gradient cryo-probe and on a Bruker Avance DRX 700 operating at a ^1H frequency of 700.13 MHz and equipped with a triple resonance, *z*-axis gradient probe. TOCSY and NOESY spectra were recorded using a mixing time of 60 ms and 150 ms, respectively, and a WATERGATE [20] pulse scheme for solvent signal suppression. Typically, 2D experiments on f5b and f6b were recorded on the 500 MHz equipped with the cryo-probe, acquiring 16 scans (64 for the NOESY), 256 experiments in the t_1 dimension, and 4 k complex points. 1D experiments on f5b and f6b were recorded with 256 scans and 16 k complex points. For f5a and f6a, 1D experiments only were acquired on the 700 MHz, typically with 2048 scans and 16 k points. Amide temperature coefficients were calculated from 2D TOCSY and 1D spectra recorded between 298 K and 302 K with a 1 K step. Additional spectra were recorded at 308, 313, and 318 K. Data were transformed using Xwin-NMR (Bruker BioSpin) and analyzed using XEASY [21]. Chemical shifts were referenced to sodium trimethylsilylpropionate.

CD spectroscopy

Samples for CD spectroscopy were prepared dissolving the lyophilized peptides in water. Peptide concentration was determined by amino acid analysis. Briefly, hydrolysis was carried out for 60 min *in vacuo* at 150 °C in the presence of 6 M HCl containing 2% phenol (w/w). Derivatization of the amino acid mixture with phenylisothiocyanate was achieved according to the standard protocol of PicoTag system (Waters). Analysis of free amino acids was performed by RP-HPLC on a PicoTag 3.9×300 mm column. The resulting peptide concentrations were: 47 μM (EGF-14), 9 μM (f5a), 51 μM (f5b), 17 μM (f6a), 51 μM (f6b). CD spectra were recorded on a Jasco J-810 spectropolarimeter using 0.1 cm and 1 cm quartz cuvettes. CD spectra of f5b and f6b were recorded between 250 and 190 nm (0.1 cm cuvette) and between 350 and 250 nm (1 cm cuvette); CD spectra of f5a and f6a were recorded between 250 and 190 nm in a 1-cm cuvette. CD spectra of native EGF-14 were recorded between 250 and 190 nm using a 0.1-cm cuvette and between 350 and 250 nm with the same path length and a 5X solution. For each sample, five scans were acquired, the baseline subtracted from the raw spectra, and the mean residue ellipticity (MRE, deg·cm^2·dmol^{-1}) was calculated dividing the CD signal intensity (mdeg) by $10 \times c \times l \times N$, where c is the peptide concentration (M), l the path length (cm), and N the number of residues.

A quantitative estimation of secondary structure content was carried out using different methods: SELCON3 [22], CONTINLL [23], CDSSTR [24,25], and K2D [26]. SELCON3, CONTINLL, and CDSSTR were run from the DichroWeb web server [27], K2D from the K2D web server [26]. SELCON3, CONTINLL, and CDSSTR were applied using a reference data set of 49 proteins, including five denatured proteins, with a wavelength range of 240–190 nm [28]. K2D does not require any reference data set and makes use of data between 240 and 200 nm only.

Results

Peptide synthesis

All peptides were prepared by standard solid-phase Fmoc-based methods, either in automatic or manual fashion. After cleavage/deprotection of the peptide-resin, the identity of the peptides was checked by LC-MS (Table 1) and the yield and purity estimated from RP-HPLC (Table 1). While products of Cys racemization were not observed, the most common side products were des-Cys-peptides and piperidide-derivate peptides. The extent of the latter modification was minimized (< 5%, as estimated by LC-MS) using 1-hydroxybenzotriazole in dimethylformamide during Fmoc deprotection and keeping deprotection times to a minimum. Deprotected, reduced peptides were purified by RP-HPLC to obtain final yields in the range 21–57% (Table 1) and purity > 99%.

Oxidative folding

The purified, lyophilized peptides were refolded in the presence of the Cys/cystine redox couple. The time course of the refolding kinetics was followed both by LC-MS and RP-HPLC. LC-MS was used to monitor the formation of disulfide bonds from the loss of two atomic mass units in molecular mass for each disulfide formed, while RP-HPLC was used to measure retention times and quantify the decrease of the starting product by UV detection at 214 nm. The reduced peptides convert rapidly in a mixture of one- and two-disulfide products (Fig. 2), which undergo a slower oxidation and reshuffling to give several three-disulfide isomers in frame-shifted peptides, or a largely predominant

Table 1. Peptide synthesis. Yield (%) of the crude deprotected peptide as estimated by weight; purity (%) of the crude deprotected peptide as estimated by RP-HPLC; final yield (%); expected and observed average molecular mass (Da) of the reduced, purified peptide.

Peptide	Yield (%)	Purity (%)	Final yield (%)	Expected mass (Da)	Observed mass (Da)
EGF-14	85.5	66.4	56.7	3451.2	3452.0
f2	84.6	46.5	39.3	3483.4	3484.3
f3	86.1	44.1	38.0	3398.3	3398.0
f4	86.5	58.6	50.7	3327.3	3328.0
f5	83.4	48.3	40.3	3477.3	3478.7
f6	83.1	25.0	20.7	3451.3	3451.5

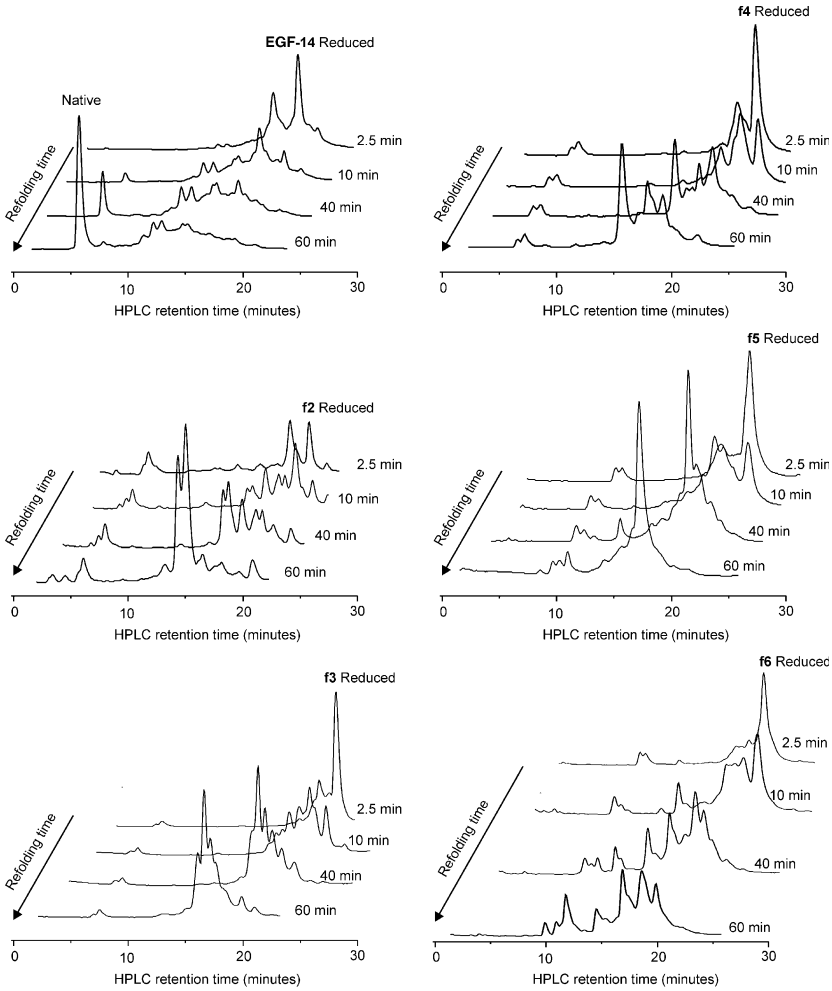

Fig. 2. Oxidative folding. RP-HPLC profiles of oxidative folding reactions, as detected by UV at 214 nm, of the different peptides at selected refolding times. The peak corresponding to the initial, fully reduced form is marked by the name of the peptide.

product for EGF-14. Under our refolding conditions, EGF-14 was rapidly oxidized and in 2 h transformed into the native three-disulfide species. On the contrary, the oxidative folding of frame-shifted peptides f2–f6 resulted in a complex mixture of oxidized isomers in all cases. The equilibrium pattern was reached within 24 h (Fig. 3) and after this time changes in the relative abundance of the species or formation of new products were not observed. The LC-MS analysis confirmed that all products in the final mixtures are three-disulfide isomers.

The quantitative analysis of the RP-HPLC profiles showed that the rates of disappearance of the reduced forms are similar but not identical (Fig. 4A). A fit of experimental data with a three-parameter negative exponential curve ($R > 0.99$, data not shown) gave an apparent rate constant value of 0.54 min^{-1} for EGF-14, and values in the range 0.22–0.26 min^{-1} for f3–f6; the fit for f2 was less good, but still gave a value (0.4 min^{-1}) that is slightly smaller than that obtained for EGF-14. A noteworthy difference in the rate of formation of three-disulfide peptides was also observed (Fig. 4B). EGF-14 reached its fully oxidized form faster than the other peptides as demonstrated by LC-MS and RP-HPLC analysis (Figs 2 and 4).

A further difference between EGF-14 and the frame-shifted peptides is represented by the change in the RP-HPLC retention time going from the reduced to the oxidized species. The final product of EGF-14 oxidative folding has a retention time that is considerably shorter with respect to the reduced species (reduced form, 21.6 min; oxidized form, 7.7 min) (Table 2), while for frame-shifted peptides most products show retention time values only slightly shorter than that of the corresponding reduced peptide. Only in the case of f5 and f6, the retention time of one of the final products is significantly reduced compared with the starting product. To quantitatively compare the behavior of the different peptides, the chromatographic parameter α, defined as the ratio between the retention time of the oxidized product (RT_{ox}) and the retention time of the reduced peptide (RT_{red}) was chosen. As shown in Table 2, EGF-14 displays the lowest α value.

EGF-14 disulfide topology

The determination of the disulfide bond topology was addressed with the peptide mapping methodology tailored on the peptide sequence and potential topology of disulfide bonds. EGF-14 was digested first by thermolysin. From the digestion two peptides were obtained, with molecular mass of 1403 and 2097 Da, respectively. The former product confirms the disulfide bridge between C611 and C620. The 2097 Da peptide, on the other hand, could not give an unequivocal answer about the two remaining bridges, which

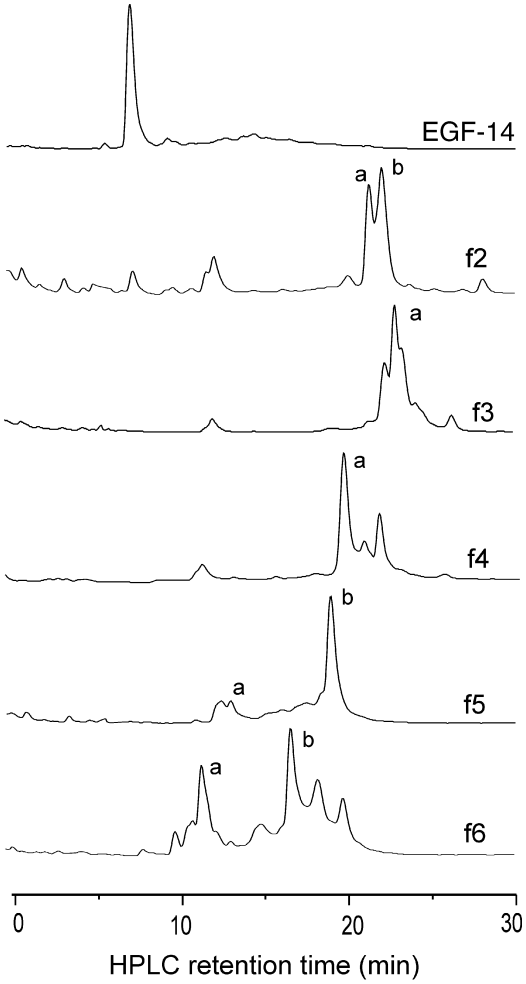

Fig. 3. Equilibrium mixtures. RP-HPLC profiles of oxidative folding reactions, as detected by UV at 214 nm, of the different peptides after 24 h. Retention times of labeled species are reported in Table 2.

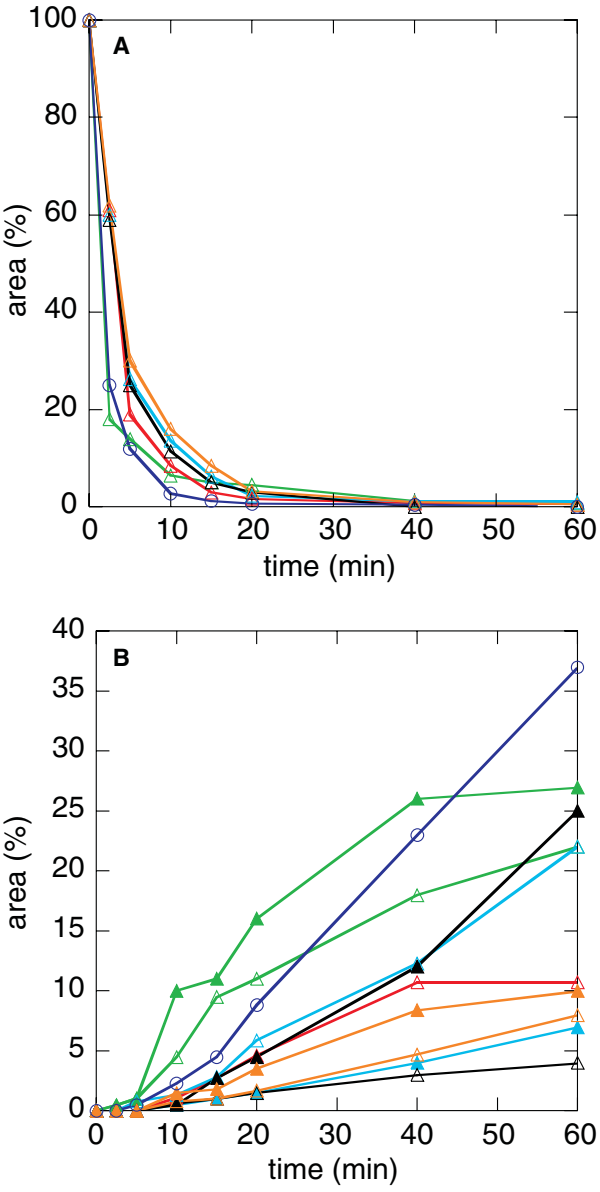

Fig. 4. Oxidative folding kinetics. (A) Disappearance of the starting product (%, area of the initial reduced form with respect to the total integrated area) for EGF-14 (blue), f2 (green), f3 (red), f4 (light blue), f5 (black), f6 (orange). (B) Formation of three-disulfide species (%, area of a three-disulfide species with respect to the total integrated area) for EGF-14 (blue), f2 (green), f3 (red), f4 (light blue), f5 (black), f6 (orange); different species (a, b in Fig. 3) originating from the same peptide are shown as empty and filled triangles, respectively. Oxidative folding kinetics were followed by RP-HPLC and UV detection at 214 nm.

could be either C594–C604/C598–C609 or C594–C609/C604–C609. The 2097 Da peptide was therefore treated with AspN endopeptidase. The reaction gave two fragments of 810 and 985 Da, respectively, which can be produced by the AspN cleavage at D597 only in the case of a C594–C604/C598–C609 combination (Table 3). The experiment thus confirms that EGF-14 from human tenascin has a disulfide topology typical of EGF domains (1–3, 2–4, 5–6).

NMR

^{1}H NMR spectra of the peptides are reported in Fig. 5, and show the drastically different dispersion in the backbone amide chemical shifts of EGF-14, in respect to that of the frame-shifted peptides. TOCSY spectra of f5b and f6b (Fig. 6) were recorded at different temperatures between 298 K and 318 K. The chemical shift dispersion of the backbone NH goups did not change in this temperature range (7.8–8.8 p.p.m. at 318 K, 7.7–8.7 p.p.m. at 298 K for f5b; 7.6–8.6 p.p.m. at 318 K, 7.7–8.8 p.p.m. at 298 K for f6b), and neither did the dispersion in the Cα chemical shifts, but the appearance of TOCSY spectra in the NH/aliphatic region is different. At 318 K, strong and sharp cross-peaks are present in the fingerprint region (3.5–5.0 p.p.m.) of f5b and, although for some residues the magnetization transfer along the side chain was not very efficient, the number of identified spin systems corresponds to the expected value. On the contrary, several cross-peaks are undetectable or have very low intensity at 298 K. After a tentative assignment of spin systems, the distribution of NH chemical shifts was compared with that expected for a random coil

Table 2. RP-HPLC retention times. Retention times of the reduced (RT_{red}, min) peptides and of the main three-disulfide species (RT_{ox}, min); difference in retention times of the reduced and oxidized forms (ΔRT, min) and selectivity parameter (α, defined as RT_{ox}/RT_{red}) for three-disulfide species.

Peptide	RT_{red} (min)	RT_{ox} (min)	ΔRT (min)	α
EGF-14	21.6	7.7	13.9	0.36
f2	28.7	f2a 21.6	7.1	0.75
		f2b 22.3	6.4	0.78
f3	27.7	23.0	4.7	0.83
f4	26.3	20.7	5.6	0.79
f5	22.8	f5a 12.3	10.5	0.54
		f5b 18.2	4.6	0.80
f6	23.2	f6a 12.6	10.6	0.54
		f6b 18.0	5.2	0.78

peptide of the same sequence [29], and with that of the native EGF-14 plotting the percentage of NH peaks in each 0.1 p.p.m. chemical shift interval (Fig. 7). The chemical shift dispersion of backbone NHs in f5b ($\sigma = 0.27$) is two times larger than that expected for a random coil peptide of the same sequence ($\sigma = 0.12$), but less than half of that of the native EGF-14 ($\sigma = 0.63$). In a similar way, the chemical shift dispersion of backbone NHs in f6b ($\sigma = 0.26$) is three times larger than that expected for a random coil peptide of the same sequence ($\sigma = 0.064$), but considerably smaller than that of the native EGF-14 ($\sigma = 0.63$). Peptide f5a showed an even smaller dispersion in the backbone NH chemical shifts compared with f5b, with broad unresolved lines in the range 7.8–8.7 p.p.m. On the contrary, f6a displayed a slightly larger chemical shift dispersion than f6b, with most of the peaks clustered in the region 7.8–8.9 p.p.m., but three NH resonances shifted downfield at 9.2–9.3 p.p.m. In a similar way, also in the methyl region, a slightly larger chemical shift dispersion was observed.

The amide NH temperature coefficients were measured between 298 K and 302 K. Such a small temperature interval was chosen to limit chemical shift variations due to temperature-induced conformational changes, and is nevertheless sufficient to measure temperature coefficients in a reliable way for most of the detectable spin systems. Measured values were more negative than -4.3 p.p.b.$\cdot K^{-1}$ and -4.8 p.p.b.$\cdot K^{-1}$ for f5b and f6b, respectively, suggesting that no stable H-bond involved in secondary structure elements is formed [30]. However, several NH amides had values in the borderline region around -4.5 p.p.b.$\cdot K^{-1}$.

The chemical shift of the aromatic protons in the three histidine residues was also compared. In f5b, the 4H protons all resonate between 7.05 and 7.10 p.p.m., while the 2H protons are well separated and resonate at 8.08, 8.27, and 8.40 p.p.m. at 298 K. In f6b, the 4H protons resonate between 7.10 and 7.15 p.p.m., and the 2H protons, which are not as well separated as in f5b, resonate between 8.32 and 8.39 p.p.m. at 298 K.

NOESY spectra of both f5b and f6b displayed very few cross-peaks, suggesting a correlation time for the molecules close to the zero-point of the NOE at that field.

CD

The CD spectrum of EGF-14 is dominated by a negative band in the far-UV region (Fig. 8A). This band has its minimum at 200 nm, a shoulder at 215 nm and is going to zero at $\approx$190 nm. Two additional much weaker positive bands can be observed in the far-UV at $\approx$235 nm and in the near-UV at 270 nm (Fig. 8B). The CD spectra of f5b and f6b (Fig. 8) are also dominated by the negative band at 200 nm and resemble that of EGF-14, but the shoulder at 215 nm and the positive bands are missing; on the contrary, the CD spectrum of f6b is slightly negative at $\approx$270 nm, and the intensity of this band is roughly four times weaker than that of EGF-14. The CD spectra of f5a and f6a could be recorded only in the far-UV region. Peptide f5a has two very weak negative bands at 205 and 230 nm, while the spectrum of f6a is characterized by a weak negative band shifted at 215 nm.

The positive CD band in the spectrum of EGF-14 in the 250–300 nm region can arise both from the contribution of the only Tyr present and from the disulfide bonds. Peptides f5b and f6b do not contain any Tyr but one Phe instead, which does not contribute significantly to the adsorption beyond 270 nm. The weak negative band displayed by f6b in this region might then arise from a partial order in the disulfide bonds. On the contrary, f5b does not show any optical activity in this range, suggesting that the disulfides are flexible.

Table 3. EGF-14 disulfide mapping. Determination of disulfide bond topology of EGF-14 by proteolysis and identification of the fragments by LC-MS. Cleavage sites are identified by a slash (/), Cys residues in bold. The disulfide pattern numbering refers to the consecutive positions of cysteines within the sequence.

Enzyme	Fragments	Mass, found (Da)	Mass, calculated (Da)	Disulfide pattern
Thermolysin	GQH**S**C**P**S**D**C**N**N(590–600)/LGQ**C**(601–604)/VSGR**C**(605–609)	2097.5 $(M+H)^{1+}$	2097.3	1–3; 2–4
		1049.8 $(M+2H)^{2+}$	1048.6	
	I**C**NEGYSGED**C**SE(610–622)	1403.8 $(M+H)^{1+}$	1403.4	5–6
		702.5 $(M+2H)^{2+}$	701.7	
	I**C**NEG(610–614)/YSGED**C**SE(615–622)	1422.0 $(M+H)^{1+}$	1421.3	5–6
		711.5 $(M+2H)^{2+}$	710.6	
aspN	S**C**P**S**(593–596)/LGQ**C**(601–604)	810.1 $(M+H)^{1+}$	811.5	1–3
		406.1 $(M+2H)^{2+}$	405.9	
	D**C**NN(597–600)/VSGR**C**(605–609)	985.5 $(M+H)^{1+}$	985.1	2–4
		493.5 $(M+2H)^{2+}$	493.1	

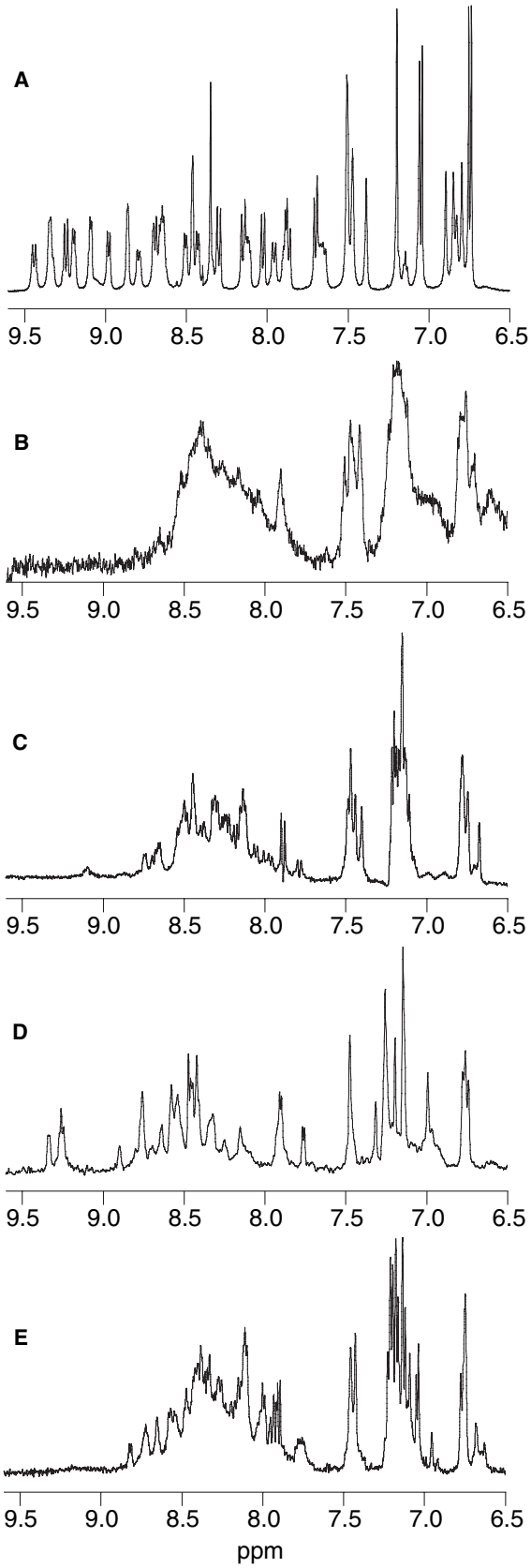

Fig. 5. NMR spectroscopy. ^{1}H-1D spectra (amide/aromatic region) of EGF-14 (A), f5a (B), f5b (C), f6a (D), f6b (E) at 298 K in H_2O/D_2O (90 : 10, v/v).

The positive band at 230 nm in the far-UV CD spectrum of EGF-14 can also arise from the contribution of Tyr. This band is not present in the spectra of the frame-shifted peptides. The other bands in this region are mainly dictated by the electronic transitions of the backbone chromophores and are sensitive to the presence of secondary structure elements. A qualitative analysis of the spectra suggests the absence of helical structure, and a dominant component of irregular structure in all the peptides.

A quantitative analysis of secondary structure content was carried out using different methods [26,27] (SELCON3 [22], CONTINLL [23], CDSSTR [24,25], K2D [26]). These CD spectra analysis programs did not produce satisfactory results in all cases. This is not surprising, given that in such small, disulfide-rich peptides containing relatively little regular secondary structure, the contribution of side chains to the overall CD spectrum can be significant. The amounts of β sheet, turn, and unordered structure found by these methods are in the range 25–35%, 15–20%, 40–65%, respectively, with no or negligible amounts of α-helix (data not shown).

Discussion

The 'frame-shift' approach

Proteins targeted to the extracellular environment can contain several tandem cysteine-rich domains [31], and the correct pairing of cysteines to form disulfide bridges is critical to reach the final native fold. In principle, two different factors can determine the pairing of cysteines to give disulfide bonds in multidomain proteins: the topology of the disulfides within each repeat, and the frame along which this topology is repeated over the amino acid chain. Human tenascin contains 14 EGF-like repeats [7,8], for a total of 84 cysteines that need to be correctly paired to form, within each repeat, the 1–3, 2–4, 5–6 disulfide bond pattern that is characteristic of EGF modules. To look into the factors that drive the consecutive modules to fold within this unique correct structural frame, we devised a simple model system that could be studied in detail by physico-chemical methods. In this approach, six peptides were selected using a window that corresponds to the average length of tenascin EGF repeats (Fig. 1). Sliding this window over the sequence of tenascin EGF repeats 13 and 14 (residues 560–622) by one cysteine at each step, we obtained six peptides that are all 33 residue long, contain six cysteines, and bear a partial overlap in the sequence. While the first peptide corresponds to the native EGF-14 repeat, the others are frame-shifted EGF repeats displaying a different pattern in the cysteine spacing. The oxidative folding of frame-shifted peptides simulates, in a way, the mispairing that would occur whether inter- rather than intra-repeat disulfide bonds form. In other words, we forced misfolding to occur within short peptides that nevertheless maintain their native sequence.

Oxidative folding

Because the EGF repeat is one of the most commonly employed building block in extracellular proteins [10,12], we wondered if there might be a kinetic reason that largely

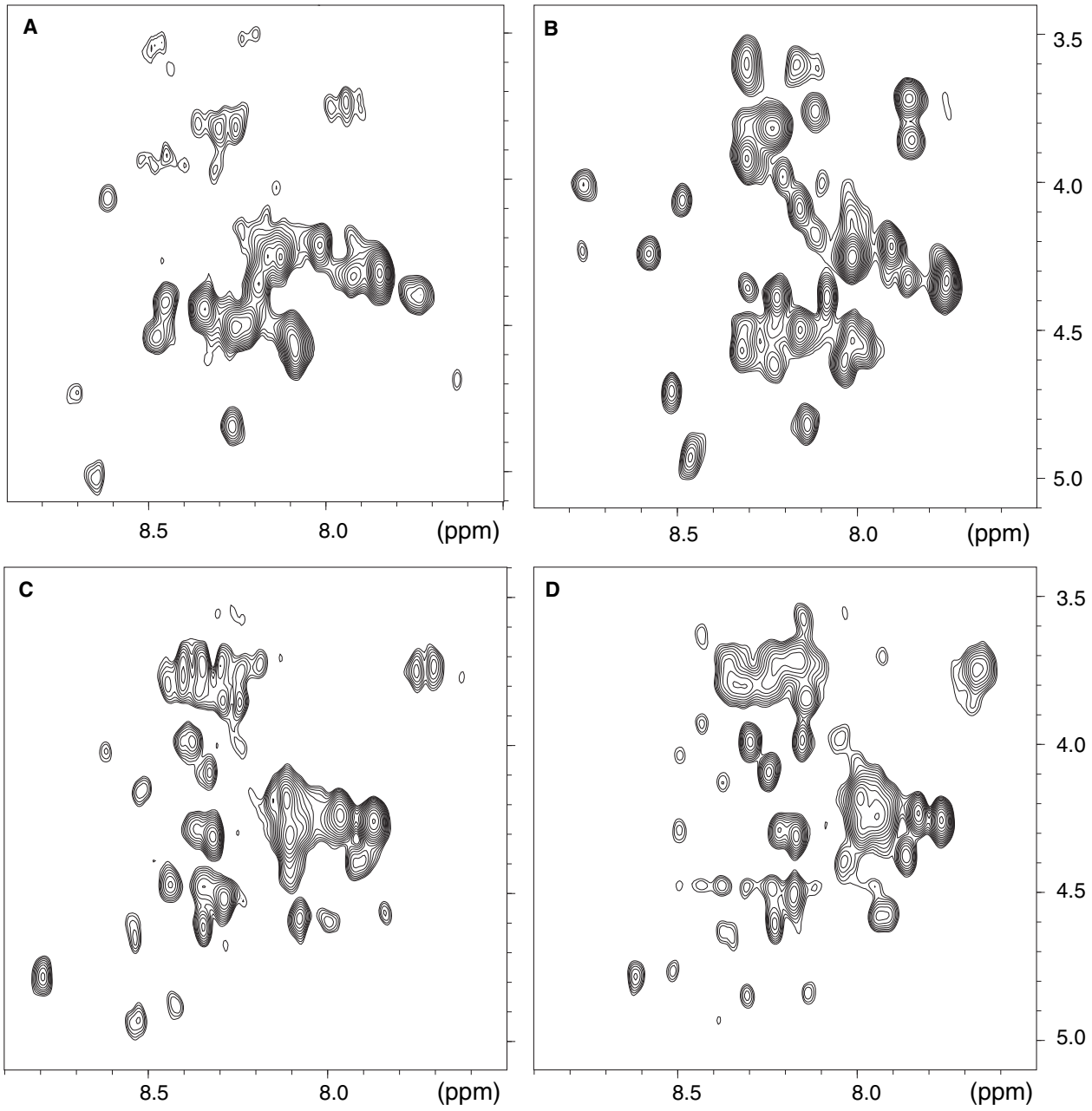

Fig. 6. NMR spectroscopy. Fingerprint region of TOCSY spectra at 298 K (left) and 318 K (right) for f5b (top, A and B) and f6b (bottom, C and D).

favors the correct formation of disulfide bonds within the same EGF repeat, or in other words, if the EGF-type repeats are so successful because they fold fast. Experimental results at least partially support this hypothesis. The disappearance rates of the reduced frame-shifted peptides, including EGF-14, are all within the same order of magnitude. The disappearance rate of the starting (reduced) product mainly reflects the oxidation rate of cysteines to cystines to form a first disulfide bond and give a species that can be separated by RP-HPLC. As the redox potential is expected to be very similar for all cysteines in the sequence in the presence of a similar chemical environment, there are no gross variations in the disappearance rate of the starting product. However, significant albeit small differences are detectable, and the oxidation of EGF-14 is slightly faster. A possible explanation is that the formation of the first disulfide in EGF-14 is favored by some residual native-like structure in the reduced state or, alternatively, that the next oxidation steps are faster, as suggested by the fact that the frame-shifted peptides only slowly evolve towards three-disulfide species, and remain trapped in a series of products, while EGF-14 is quickly finding its pathway to the native form, which within 2 h is the major species. The burial of a disulfide bond in a native-like environment can for example alter its redox potential

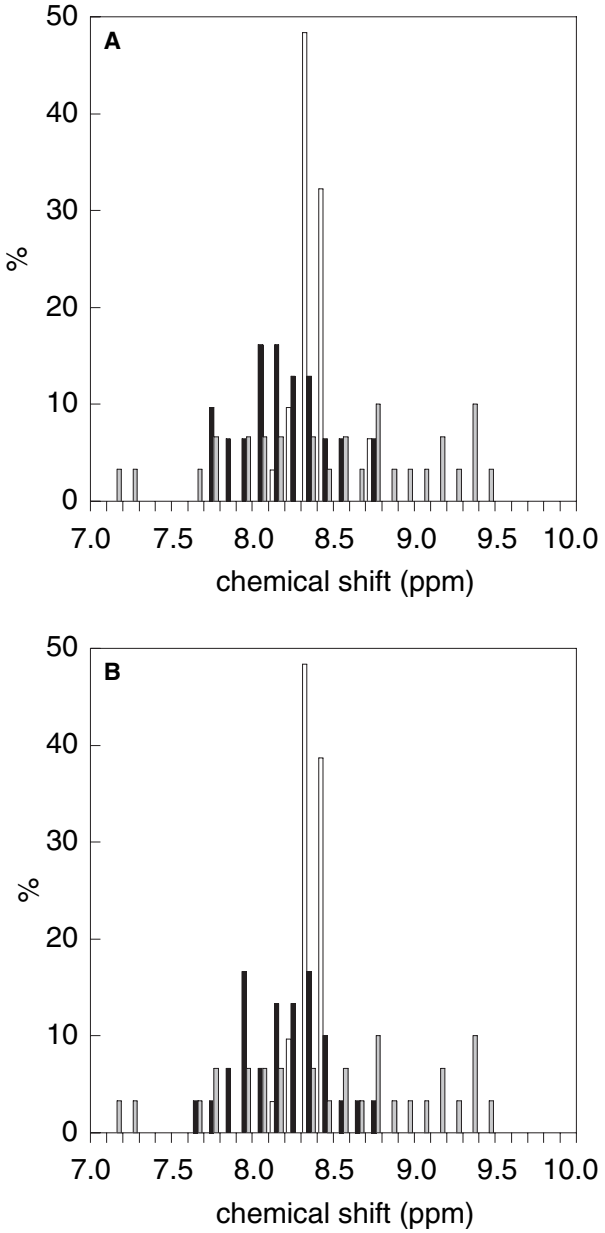

Fig. 7. Backbone NH ^{1}H chemical shifts. Distribution (%, black bars) of backbone NH chemical shifts (p.p.m.) for f5b (A) and f6b (B). The distribution of backbone NH chemical shifts of EGF-14 (gray bars) and that expected for a random coil peptide of the same sequence (white bars) are also shown.

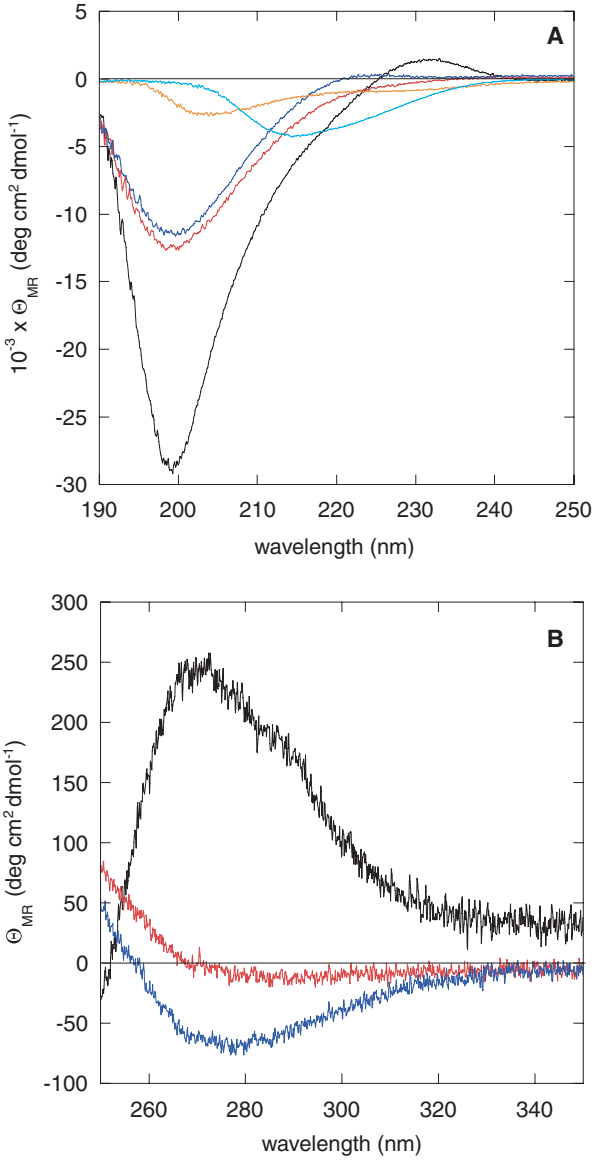

Fig. 8. CD spectroscopy. CD spectra (mean residue ellipticity, Θ_{MR}, deg·cm^2·dmol^{-1}) in the far-UV (190–250 nm, A) of EGF-14 (black), f5b (red), f6b (blue), f5a (orange) and f6a (light blue) and in the near-UV (250–350 nm, B) of EGF-14 (black), f5b (red), f6b (blue).

and render it less accessible to the external redox couple. Both kinetic (disappearance rate of the reduced peptide and convergence towards a unique product) and thermodynamic (stability of the three-disulfide species formed) factors are therefore favoring the EGF-like topology, determining a preferential 'folding frame' in the cluster of highly repeated domains.

Peptide structure

Despite the complexity of the mixtures obtained in the oxidative folding reactions, we were able to isolate and characterize, by NMR and CD, some of three-disulfide species that are formed. Our efforts were pointed towards the characterization of those products that displayed a large difference with respect to the retention time of the reduced and fully oxidized species. This was considered an important indication of effective burial of hydrophobic residues upon folding, with the formation of a relatively compact structure. This, in turn, can be promoted by a 'crossed' disulfide topology of the EGF type (1–3, 2–4, 5–6) or equivalent, while a linear arrangement of disulfides (1–2, 3–4, 5–6) is less likely to produce compact structures. NMR and CD studies suggest that the products of the oxidative folding of frame-shifted peptides (f5a, f5b, f6a, f6b) are highly flexible in solution and only partially structured, with some degree of conformational restraint given by the presence of three

disulfide bonds. In contrast, EGF-14 displays the dispersion in the NH chemical shifts and the CD characteristics of a compact globular domain.

Relevance to folding *in vivo*

The folding *in vivo* of an extracellular protein containing disulfide-rich domains, the low-density lipoprotein receptor, has shed new light on the folding process in the living cell [32]. In contrast to the commonly assumed 'vectorial' model, in which domains in a multidomain protein would fold independently and sequentially from the N- to the C-terminus, a different scenario has been proposed. In this view, after the initial polypeptide chain collapse leading to the formation of non-native disulfide bonds that can be formed even between distant cysteines, an extensive reshuffling of disulfide bonds occurs, in a rate-limiting process that is eventually leading to the native structure. Therefore, folding would mainly be a post-translational event. Reshuffling of non-native disulfide bonds, on the other hand, is carried out by the protein disulfide isomerase enzymes [33], which operate in concerted action and in physical association with chaperone proteins [34]. The mechanism through which a polypeptide chain is recognized as misfolded by the protein disulfide isomerases is not known in detail yet. The structure of an entire protein disulfide isomerase is still lacking, but homology modeling of the peptide recognition domain b′ of protein disulfide isomerase [35] suggests that a small hydrophobic pocket capable of hosting even single amino acids could represent the binding site. In a similar way, an heptapeptide fragment of alternating hydrophobic residues has been shown to be recognized by BiP [36], a mammalian chaperone of the HSP70 family. Because the primary quality control system in charge of rearranging a misfolded polypeptide chain in the lumen of the endoplasmic reticulum must be relatively unspecific in terms of amino acid sequence and secondary structure recognition, the exposure of hydrophobic residues to the solvent is the simple structural feature that might drive the reshuffling of disulfide bonds *in vivo*. There is also strong evidence that the higher the stability of the folded protein, the higher the secretion level [37], which suggests that the dynamic behavior of the polypeptide chain in the folding/unfolding process can direct it either to secretion or to degradation.

Some analogy between the folding *in vivo* and the oxidative folding of our model peptides derived from the tenascin sequence can be drawn. While in principle all possible combinations of cysteine pairing are possible in the native polypeptide chain, as shown by the fact that also frame-shifted peptides eventually evolve towards three-disulfide species, both a kinetic and a thermodynamic selection is taking place during the oxidative folding process. The kinetic selection is acting at the level of the disappearance of the starting reduced peptide, which is slightly faster for the native EGF-14. The slow step remains, however, the reshuffling of disulfide bonds. During this step, the thermodynamic selection is acting to reach, when possible, a compact, globular structure. This is the case for EGF-14, but not for the frame-shifted peptides, which exhibit only a partially folded, flexible structure. What marks the border between the properly folded native EGF-14 and the partially folded frame-shifted peptides is the less effective burial of hydrophobic residues in the latter, as evidenced by the difference in retention time between the reduced and oxidized form, which is highest in the native EGF-14. This is apparently the same mechanism underlying the recognition of a misfolded polypeptide by chaperone proteins, and probably by protein disulfide isomerases.

The experimental development of an *in vivo* model for the folding of a relatively small and well defined molecular system such the one studied here *in vitro* would enable us to run parallel studies with significant outcomes in the comprehension of the oxidative folding of disulfide-rich modular proteins in eukaryotic cells.

Acknowledgement

We acknowledge the support of the European Community – Improving Human Potential Programme – Access to Research Infrastructures for use of the NMR spectrometers at the PARABIO Large Scale Facility (CERM) in Sesto Fiorentino, Italy. We are grateful to Doriano Lamba (CNR) for use of the circular dichroism spectropolarimeter, to Roberta Pierattelli (CERM), Massimo Lucci (CERM) and Gennaro Esposito (University of Udine) for assistance with NMR data acquisition, and to Sotir Zahariev (ICGEB) for helpful suggestions. This work is part of F.Z. PhD thesis at ICGEB-SISSA.

References

1. Mackie, E.J. (1997) Molecules in focus: tenascin-C. *Int. J. Biochem. Cell. Biol.* **29**, 1133–1137.
2. Jones, P.L. & Jones, F.S. (2000) Tenascin-C in development and disease: gene regulation and cell function. *Matrix Biol.* **19**, 581–596.
3. Chiquet-Ehrismann, R. & Chiquet, M. (2003) Tenascins: regulation and putative functions during pathological stress. *J. Pathol.* **200**, 488–499.
4. Ohta, M., Sakai, T., Saga, Y., Aizawa, S. & Saito, M. (1998) Suppression of hematopoietic activity in tenascin-C-deficient mice. *Blood* **91**, 4074–4083.
5. Mackie, E.J. & Tucker, R.P. (1999) The tenascin-C knockout revisited. *J. Cell. Sci.* **112**, 3847–3853.
6. Jones, F.S., Hoffman, S., Cunningham, B.A. & Edelman, G.M. (1989) A detailed structural model of cytotactin: protein homologies, alternative RNA splicing, and binding regions. *Proc. Natl Acad. Sci. USA* **86**, 1905–1909.
7. Gulcher, J.R., Nies, D.E., Alexakos, M.J., Ravikant, N.A., Sturgill, M.E., Marton, L.S. & Stefansson, K. (1991) Structure of the human hexabrachion (tenascin) gene. *Proc. Natl Acad. Sci. USA* **88**, 9438–9442.
8. Nies, D.E., Hemesath, T.J., Kim, J.H., Gulcher, J.R. & Stefansson, K. (1991) The complete cDNA sequence of human hexabrachion (Tenascin): a multidomain protein containing unique epidermal growth factor repeats. *J. Biol. Chem.* **266**, 2818–2823.
9. Swindle, C.S., Tran, K.T., Johnson, T.D., Banerjee, P., Mayes, A.M., Griffith, L. & Wells, A. (2001) Epidermal growth factor (EGF)-like repeats of human tenascin-C as ligands for EGF receptor. *J. Cell. Biol.* **154**, 459–468.
10. Campbell, I.D. & Bork, P. (1993) Epidermal growth factor-like modules. *Curr. Opin. Struct. Biol.* **3**, 385–392.
11. Hommel, U., Harvey, T.S., Driscoll, P.C. & Campbell, I.D. (1992) Human epidermal growth factor: high resolution solution structure and comparison with human transforming growth factor alpha. *J. Mol. Biol.* **227**, 271–282.

12. Letunic, I., Copley, R.R., Schmidt, S., Ciccarelli, F.D., Doerks, T., Schultz, J., Ponting, C.P. & Bork, P. (2004) SMART 4.0: towards genomic data integration. *Nucleic Acids Res.* **32**, D142–D144.
13. Chang, J.Y., Li, L. & Lai, P.H. (2001) A major kinetic trap for the oxidative folding of human epidermal growth factor. *J. Biol. Chem.* **276**, 4845–4852.
14. Chang, J.Y., Li, L. & Bulychev, A. (2000) The underlying mechanism for the diversity of disulfide folding pathways. *J. Biol. Chem.* **275**, 8287–8289.
15. Chang, J.Y. & Li, L. (2002) The disulfide structure of denatured epidermal growth factor: preparation of scrambled disulfide isomers. *J. Protein Chem.* **21**, 203–213.
16. Barnham, K.J., Torres, A.M., Alewood, D., Alewood, P.F., Domagala, T., Nice, E.C. & Norton, R.S. (1998) Role of the 6–20 disulfide bridge in the structure and activity of epidermal growth factor. *Protein Sci.* **7**, 1738–1749.
17. Valcarce, C., Bjork, I. & Stenflo, J. (1999) The epidermal growth factor precursor: a calcium-binding, beta-hydroxyasparagine containing modular protein present on the surface of platelets. *Eur. J. Biochem.* **260**, 200–207.
18. Han, Y., Albericio, F. & Barany, G. (1997) Occurrence and minimization of cysteine racemization during stepwise solid-phase peptide synthesis. *J. Org. Chem.* **62**, 4307–4312.
19. Kaiser, E., Colescott, R.L., Bossinger, C.D. & Cooke, P.I. (1970) Color test for detection of free terminal amino groups in the solid-phase synthesis of peptides. *Anal. Biochem.* **34**, 595–598.
20. Piotto, M., Saudek, V. & Sklenar, V. (1992) Gradient-tailored excitation for single-quantum NMR spectroscopy of aqueous solutions. *J. Biomol. NMR* **2**, 661–665.
21. Bartels, C., Xia, T.-H., Billeter, M., Güntert, P. & Wüthrich, K. (1995) The program XEASY for computer-supported NMR spectral analysis of biological macromolecules. *J. Biomol. NMR* **5**, 1–10.
22. Sreerama, N. & Woody, R.W. (1993) A self-consistent method for the analysis of protein secondary structure from circular dichroism. *Anal. Biochem.* **209**, 32–44.
23. van Stokkum, I.H., Spoelder, H.J., Bloemendal, M., van Grondelle, R. & Groen, F.C. (1990) Estimation of protein secondary structure and error analysis from circular dichroism spectra. *Anal. Biochem.* **191**, 110–118.
24. Manavalan, P. & Johnson, W.C. Jr (1987) Variable selection method improves the prediction of protein secondary structure from circular dichroism spectra. *Anal. Biochem.* **167**, 76–85.
25. Sreerama, N. & Woody, R.W. (2000) Estimation of protein secondary structure from circular dichroism spectra: comparison of CONTIN, SELCON, and CDSSTR methods with an expanded reference set. *Anal. Biochem.* **287**, 252–260.
26. Andrade, M.A., Chacon, P., Merelo, J.J. & Moran, F. (1993) Evaluation of secondary structure of proteins from UV circular dichroism spectra using an unsupervised learning neural network. *Protein Eng.* **6**, 383–390.
27. Lobley, A., Whitmore, L. & Wallace, B.A. (2002) DICHROWEB: an interactive website for the analysis of protein secondary structure from circular dichroism spectra. *Bioinformatics* **18**, 211–212.
28. Sreerama, N., Venyaminov, S.Y. & Woody, R.W. (2000) Estimation of protein secondary structure from circular dichroism spectra: inclusion of denatured proteins with native proteins in the analysis. *Anal. Biochem.* **287**, 243–251.
29. Wüthrich, K. (1986) *NMR of Proteins and Nucleic Acids*. John Wiley & Sons, New York.
30. Cierpicki, T. & Otlewski, J. (2001) Amide proton temperature coefficients as hydrogen bond indicators in proteins. *J. Biomol. NMR* **21**, 249–261.
31. Bork, P., Downing, A.K., Kieffer, B. & Campbell, I.D. (1996) Structure and distribution of modules in extracellular proteins. *Q. Rev. Biophys.* **29**, 119–167.
32. Jansens, A., van Duijn, E. & Braakman, I. (2002) Coordinated nonvectorial folding in a newly synthesized multidomain protein. *Science* **298**, 2401–2403.
33. Freedman, R.B., Klappa, P. & Ruddock, L.W. (2002) Protein disulfide isomerases exploit synergy between catalytic and specific binding domains. *EMBO Rep.* **3**, 136–140.
34. Ellgaard, L. & Helenius, A. (2003) Quality control in the endoplasmic reticulum. *Nat. Rev. Mol. Cell. Biol.* **4**, 181–191.
35. Pirneskoski, A., Klappa, P., Lobell, M., Williamson, R.A., Byrne, L., Alanen, H.I., Salo, K.E., Kivirikko, K.I., Freedman, R.B. & Ruddock, L.W. (2004) Molecular characterization of the principal substrate binding site of the ubiquitous folding catalyst protein disulfide isomerase. *J. Biol. Chem.* **279**, 10374–10381.
36. Blond-Elguindi, S., Cwirla, S.E., Dower, W.J., Lipshutz, R.J., Sprang, S.R., Sambrook, J.F. & Gething, M.J. (1993) Affinity panning of a library of peptides displayed on bacteriophages reveals the binding specificity of BiP. *Cell* **75**, 717–728.
37. Kowalski, J.M., Parekh, R.N., Mao, J. & Wittrup, K.D. (1998) Protein folding stability can determine the efficiency of escape from endoplasmic reticulum quality control. *J. Biol. Chem.* **273**, 19453–19458.
38. Kelley, L.A., MacCallum, R.M. & Sternberg, M.J.E. (2000) Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J. Mol. Biol.* **299**, 499–520.
39. Sali, A. & Blundell, T.L. (1993) Comparative protein modeling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779–815.