Scuola Internazionale Superiore di Studi Avanzati - Trieste

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

Thesis submitted for the Degree of Doctor Philosophiae.

Candidate:

Francesco Zanuttin

Supervisor:

Prof. Sándor Pongor

SISSA – Via Beirut 2-4 – 34014 TRIESTE – ITALY

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

Thesis submitted for the Degree of Doctor Philosophiae.

Candidate:

Francesco Zanuttin

Supervisor:

Prof. Sándor Pongor

Acknowledgments

We have always a lot to learn from people around us, so that at the end of the day we cannot say we have done something really alone. This is why I would like to thank:

Sándor, a multiple interest scientist, for giving me the possibility to work in his group,

Corrado, for always being helpful and positive,

Alessandro, for teaching me how to make science,

Sotir, a genuine enthusiast of science: especially of.....pure chemistry,

The smiling Kristian, a very human computational biologist,

Maristella, Maša and Gordana for their help and friendship,

and of course, Aila.

Abstract

Epidermal growth factor (egf) domains are 30-50 residue long repeats characterized by the strict conservation of six cysteine residues, which are forming three disulfide bonds with the topology 1-3, 2-4, 5-6. The common structural feature of egf domains is a two-stranded beta-sheet from which the three disulfide bonds depart to connect the N- and C-terminal loops, to make a rather compact structure. 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. In order to investigate the factors that determine the correct folding of epidermal growth factor-like repeats (egf) within a multi-domain protein, we prepared a series of six peptides that, taken together, span the sequence of two egf repeats of human tenascin, a large extra-cellular matrix glycoprotein expressed during embryonic development and in proliferative processes such as wound healing and tumorigenesis. 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 residue 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 to 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*.

The results of this work are published in:

Zanuttin, F., Guarnaccia, C., Pintar, A., Pongor, S.

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

European Journal of Biochemistry. 2004 Nov; 271 (21): 4229-40.

Contents

1 Introduction

1.1 The epidermal growth factor

The epidermal growth factor (EGF) was the first growth factor to be discovered. It was isolated for the first time by Stanley Cohen and coworkers in the 1962 from a submaxillary gland extract (1). This 54 amino acids peptide belongs to a large family of molecules with the ability to modulate cellular growth. EGF has been demonstrated to elicit significant biological responses in cell culture systems as well as *in vivo*, inducing cell proliferation, particularly of keratinocytes and fibroblasts (2). In the organism the EGF production is stimulated by testosterone and inhibited by estrogens (3,4). EGF is expressed as a trans-membrane protein from which the mature peptide is released by proteolysis (5). EGF is present in milk, saliva, urine, and plasma and in most other body fluids such as pancreatic juice (2). The 6 KDa peptide isolated by Cohen is the prototype of a heterogeneous family of growth-promoting proteins, which all share one or more EGF or EGF-like structural units. A variety of different growth factors belong to the family of the EGF-like, among them the transforming growth factor-α (TGF- α) and the heparin-binding EGF-like growth factor (HB-EGF). The EGF-like unit is defined by six conserved cysteines in a 35-45 amino acid sequence. These cysteines form three disulfide bonds (6)with a typical conserved topology C1-C3, C2-C4, C5-C6 (7).

All these growth factors bind to four homologous EGF receptors (EGFR: EGFr (ErbB1/HER1), ErbB2 (HER2), ErbB3 (HER3) and ErbB4 (HER4)) (8). The structure of EGF receptors is a prototype of receptor tyrosine kinases (9,10). They are formed by a tyrosine kinase cytoplasmic domain with a C-terminal regulatory sequence, a single transmembrane lipophylic α-helix and an N-terminal extracellular domain involved in ligand binding and receptor dimerization. Like most of the protein-tyrosine kinase receptors known so far, also EGF receptors consist of a single polypeptide chain.

EGF-like growth factors bind EGFR monomers, promoting receptor dimerization and oligomerization(11). Consequently to the growth factor binding, the receptors forming homo or hetero-dimers activate each other by transphosphorylation of the cytoplasmic tails. The phosphorylated tyrosines, usually six, are docking sites for effector molecules.

These intracellular factors are generally other kinases, which trigger a complex network of intracellular signal transduction pathways that regulate cell growth.

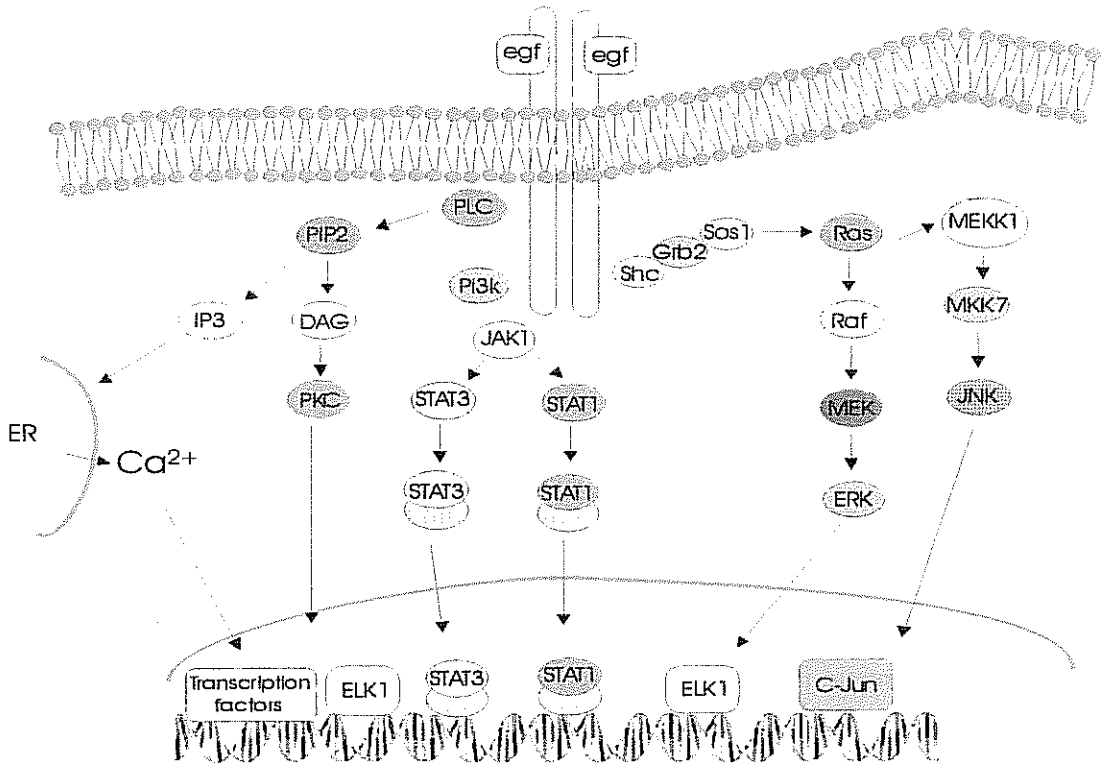


Figure 1: The proliferative effects of EGF are signaled through several pathways, Binding of EGF results in ERFR dimerization, autophosphorylation and tirosin phosphorylation of other proteins. The EGF receptor activates ras and MAP kinase pathways, ultimately causing phosphorylation of transcription factors sach as c-Fos to create AP-1 and ELK-1. Activation of STAT-1 and STAT-3 transcription factors by JAK kinases and Phosphatidylinositol signaling contributes to proliferative effect. Crosstalk of EGF signaling with other pathways makes the EGF receptor a junction point between different signaling systems.

The activated EGF receptor recruits ras (12) and MAP kinase, finally causing phosphorylation of transcription factors such as c-Fos to create AP-1 and ELK-1. STAT-1 and STAT-3 are also activated by JAK kinases in response to EGF stimulation establishing a fast-responding connection between trans-membrane receptors and transcription factors. Another component of EGF signaling is protein kinase C, which is activated by phosphatidylinositol signaling and calcium release (13). Many points of the EGF signaling are shared with other pathways, making EGF one of the elements of a complex cross-talk network regulating cell growth (14).

An increased activity of growth factors and their receptors can lead to an uncontrolled cell proliferation. It is not surprising that tumors express high levels of these proteins and an overexpression of the EGF receptor family is frequently observed in human carcinomas. An increased expression of EGFR has been demonstrated in a variety of solid tumors, particularly in breast, (15), colon, prostate (16) and ovarian cancer (17). In particular, the over-expression of two of the EGFR family receptor, EGFr and ErbB2, is associated with advanced stages of the disease, more aggressive clinical behavior and poor prognosis for the patient. The role played by EGF receptors in human cancer development makes them a possible target for new treatment strategies (18). For this reason possible therapies directed at blocking the function of these receptors have been the argument of intensive studies in the last decade (19,20).

1.2 EGF like motifs

Sequences with a significant homology to EGF are present in many different proteins either in single copy or arranged in multiple tandem repeats. EGF-like modules usually comprise from 45 to 53 residues with 6 cysteine residues linked to form three disulfide bonds. This module is present in a wide variety of multidomain proteins where it probably mediates similar biological functions. The EGF-like superfamily is formed by three families, EGF-laminin, EGF-like motifs and EGF-Calcium binding modules. The Laminin EGF family is distinguished by having a C-terminal extension with an additional disulphide bond. Among different EGF-like domains, the number and type of amino acids between cysteines can be significantly different. Considering a comparison of many different EGF-like modules, the following consensus sequence was proposed: (21).

$$xxxxxCx_{2-7}Cx_{1-4}(G/A)xCx_{1-13}ttaxCxCxxGax_{1-6}GxxCx$$

Where: A is an aromatic amino acid, t is a non-hydrophobic and x denotes any possible residue.

Some multidomain proteins of the extracellular matrix like laminin, tenascin and thrombospondin present several consecutive copies of the EGF-like repeat. Fibrillin, for example, contains 47 EGF-like modules arranged in 8 consecutive clusters while the Notch receptors, involved in the development, contain 36 EGF-like modules in tandem (22). These clusters are supposed to be originated by gene duplication and exon shuffling events. The multidomain glycoproteins are thought to possess specific biological activities in morphogenesis and cell motility, development, tissue repair (23,24), but the exact role played by EGF-like modules as integral elements of the extracellular matrix remains fairly uncertain (25,26). At least in two of these cases, laminin and tenascin-C, the EGF-like repeats have been demonstrated to trigger EGFR signaling acting as low affinity ligands or enhancers of soluble EGF (27).

EGF-like domains are also part of different trans-membrane proteins such as the low-density lipoprotein receptor (LDLR) (28) and cell surface receptors involved in

development. The three N-terminal EGF-like modules present in this protein are named EGF-A, B and C. They participate in LDL binding and internalization of the receptor.

Domains with a close or distant homology with the EGF module are present even in some proteins involved in blood clotting, complement system and neural development (29). The coagulation enzymes, factors VII, IX and X and protein C, all have two EGF-like modules (30).

Several EGF-like domains present typical post-translational modifications. These are the *O*-fucosylation, *O*-glucosylation of the serine and threonine residue and *β*-hydroxylation of asparagine and aspartic acid (31). EGF-like modules of serum glycoproteins involved in blood clotting are subject to *O*-glucosylation and *O*-fucosylation. These saccaridic residues once attached to the protein can be elongated to give *O*-glucose and *O*-fucose trisaccharides. Among all different proteins presenting EGF-like modules only few of them are subjected to glycosylation. Broad consensus sequences for different type of modifications have been identified (31).

A large subset of EGF-like domains bind one Ca^{2+} ion, which is important for the orientation of neighboring modules required for biological activity (32). Coagulation factors VII, IX and X, protein C, protein S, fibrillin-1 and Notch receptors, all contain Ca^{2+} binding EGF-like modules. The Ca^{2+} binding capacity is associated with β-hydroxylation of a particular Asp or Asn residue (33). Site-directed mutagenesis demonstrated that EGF-like modules unable to bind Ca^{2+} ion give biologically inactive proteins (34,35).

In many cases EGF modules have been found to be involved in protein-protein interactions (22,36,37). The most remarkable examples are represented by the LDL receptor, Notch receptors, coagulation factor VII and urokinase receptor. In the latter, a single EGF mediates the interaction with its ligand. The hypothesis of a biological active role of EGF-like modules is also supported by the existence of post-translational modifications that might have a regulatory function. In most cases, however, the exact biological role of the EGF-like module is still unknown (38).

1.2.1 EGF-like domain structure

The archetype of proteins that contain EGF-like modules is the EGF precursor. This large trans-membrane protein contains nine EGF modules in its amino-terminal extracellular region. The EGF module closest to the cell membrane is released as active soluble EGF after proteolytic cleavage of the precursor operated by a metalloproteinase (5).

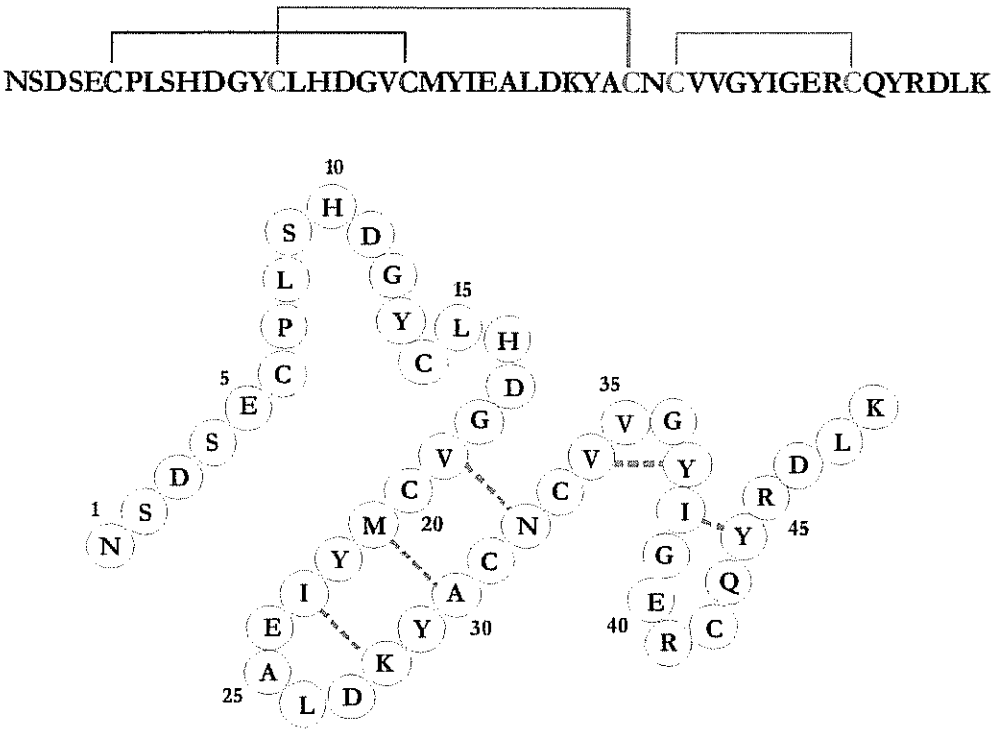


Figure 2 Schematic diagram of the secondary structural features of residues 1 to 48 of human EGF. Disulfide bonds are marked with yellow solid lines, while hydrogen bonds are indicated with dashed blue lines. Amino acids are represented by the standard one-letter code.

The fate and role of the large amino terminal fragment containing the remaining eight EGF modules after release of the soluble peptide is to date unknown. Diffusible EGF is a small protein containing 53 amino acids, its structure being stabilized by three disulfide bonds. Its six Cys residues form a conserved pattern of disulfide bonds between residues 6-20, 14-31 and 33-42, which is characteristic of all EGF modules. The human EGF (hEGF) is characterised by a two sub-domain structure; an N-terminal domain spanning from residue 1 to residue 32 and a C-terminal one from residue 38 to residue 48 (39,40).

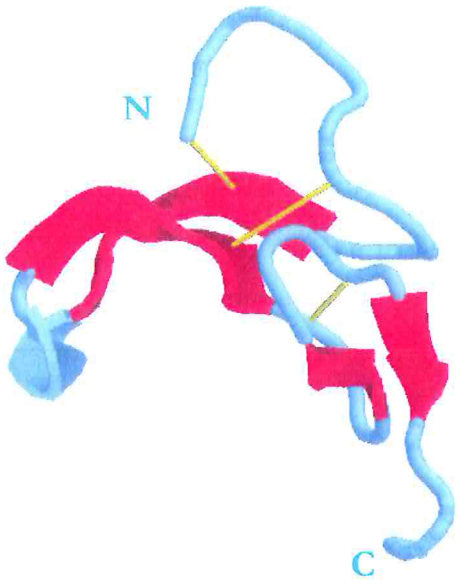


Figure 3: EGF molecule consists of two subdomains, residues from 1 to 32 and residues from33 to 48. Residues 19 to 23 and 28 to 32 form an antiparallel β-sheet that is the most important structural feature for biological activity. The C-terminal β-sheet involves residues 37, 38 and 44, 45.The first five N-terminal amino acids are not shown in this model

The two sub-domains are linked by a tight turn (Val34-Gly37). The N terminal sub-domain contains an antiparallel β-sheet (19-32) and two disulfide bonds involving Cys 6-20 and 14-31. The two strands of the β sheet are connected by a β-turn formed by residues 24 to 27. The first five N terminal residues are very mobile and unstructured, but can weakly interact with the β-sheet forming a transitory third strand. The smaller C-terminal sub-domain contains one disulfide bond and forms a minor β-sheet made up by residues 37-38 and 44-45 connected by a turn in region 40. Highly conserved aromatic residues, Tyr 13, 22 and 29 in hEGF, are close in space and form an aromatic cluster.
The minimal part of EGF that can independently bind the EGF receptor is the peptide encompassing residues from 20 to 31. This segment includes the N-terminal β-sheet domain and is considered the most important part for biological function as confirmed by the loss of activity caused by the disruption of this structure.

As the four loops formed by the three-disulfide bond constraints can be different in length, the structure of distinct EGF-like domains can present significant differences. Among all the structures the antiparallel β-sheet between Cys 20 and Cys 31 appear as the most conserved motif. The other regions show less conserved structures even if the

position of turns appears to be more regular, which probably explains the importance of the few highly conserved residues in the EGF-like consensus sequence (38). The EGF motif forms a very stable structure that, in the absence of reducing agents, can be denaturated only under harsh conditions. A study on murine EGF suggests that the disulfide bond pattern is not the determinant feature in establishing the native fold. The EGF backbone fold, in fact, might accommodate two disulfide patterns other than the native one (1-3,2-4,5-6 and 1-3, 2-5, 4-6). Both of them perfectly satisfy the structural restraints being indistinguishable from the native one. From a kinetic point of view the native pattern is reached flowing through three groups of highly heterogeneous folding intermediates. Among the five 1-disulfide isomers characterized in the folding pathway of human EGF, only one presents a native disulfide bond. All other species have non-native disulfide bonds mostly established between neighboring cysteines. A single stable 2-disulfide isomer (EGF-II) is considered the main kinetic trap during the folding reaction, representing more than 85% of the total protein. EGF-II is characterized by the presence of two native disulfides (14-31, 33-42). Nevertheless, the formation of the final bridge (6-20) is a very slow process and EGF-II can rather accomplish the native state through non-native 3-disulfide scrambled isomers. EGF and EGF-like modules present three major peculiar characteristics among other known disulfide rich proteins: (i) The compatibility of the native structure with different, even non-native, disulfide patterns, (ii) the presence of several non-native isomers along the folding pathway and (iii) the low sequence similarity within the family. Even if the folding of this peptide has been largely investigated, it is not possible up to now to draw a comprehensive general description of the process. Moreover is the driving force that leads such a heterogenic group of proteins to reach the same disulfide bound pattern and a similar foldremains unclear.

1.3 Disulfide bonds

Cysteine residues in proteins targeted to the extra-cellular environment can undergo an oxidative reaction to form disulfide bridges. Disulfide bonds constitute a stable constraint that strongly increase the protein stability, and allow proteins to survive in the extracellular environment and in harsh conditions like strong acids or boiling (47). The disulfide bond is not linear and prefers non-planar conformations introducing a peculiar asymmetry in the protein structure. The distribution of the value of the dihedral angle χ_3 shows two peaks at -80° for the left-handed disulfides and at $+100^{\circ}$ for the right-handed bridges. Conformational parameters like distance between Cα carbons and other dihedral angles are different in left-handed and right-handed disulfide bridges while the average distance of the two sulfur atoms is in average of 2.02 Å in both conformations.

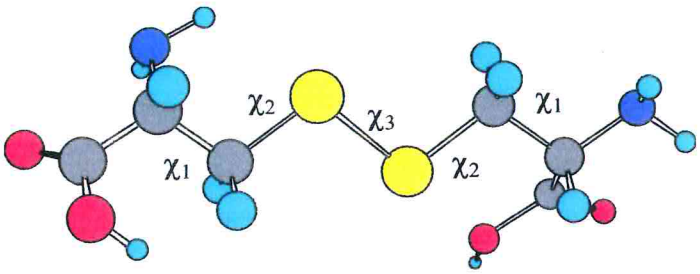


Figure 1: Structure of disulfide cross-links. Note that the two methylene groups are not in the same plane. The illustration defines the different dihedral angle in the disulfide bridge.

Cysteine residues not involved in disulfide bridges are generally deeply buried into the structure respects to the linked that are more accessible by the solvent (48). Free cysteines are preferentially present in α-elices while residues involved in disulfide bonds do not show a clear structural preference. Most of the proteins contain an even number of Cys residues, demonstrating that the disulfide bond formation is the favored condition (48). Generally proteins present only one disulfide bridge, and the average spacing between cysteines is of 11 or 16 residues. Small proteins contain a larger number of disulfide bridges with respect to larger proteins, and peptides up to 50 amino acids do not contain free cysteine residues.

Generally a protein lacking some of its disulfide bonds can maintain its native conformation even if with a different stability (49). For example, bovine pancreatic trypsin inhibitor (BPTI) adopts its native conformation even without any one of its three disulfides (50). Many other proteins show the same behavior and in some cases like the immunoglobulin domain (51) and TEM-1β-lactamase the unique disulfide bond has a very marginal role in protein folding and functionality (52).

1.3.1 Disulfide bond formation in vivo

Both in eukaryotic and prokaryotic cells, the disulfide bond formation and isomerization process takes place only in the extra-cytoplasmatic environments while in the cytosol disulfide bridges are formed only transiently as part of the catalytic cycle of enzymes. Disulfide bridge formation occurs in vivo in the endoplasmatic reticulum of eukaryotic cells and in the periplasm of Gram-negative bacteria. In these compartments disulfide bonds are very stable because the redox couple GSH/GSSG maintains a much more oxidizing environment in respect to the cytoplasm. It has been observed that even if performed in the presence of physiological concentrations of GSH/GSSG, the disulfide bond formation *in vitro* is much slower than *in vivo*. The process has been extensively studied in bacteria where, four proteins named DsbA, DsbB, DsbC and DsbD, are the crucial elements of the two distinct pathways of disulfide formation and isomerization (53). DsbA is a small periplasmic protein that belongs to the thioredoxin superfamily. The members of this family share a typical fold and contain an active site motif: -Cys_1-Gly-His-Cys_2 (54,55) (Fig. 4). The accessible thiol group of Cys_1 residue can react with other disulfide bonds as a redox agent. The product is an unstable mixed-disulfide which can spontaneously give the correct disulfide bond on the substrate protein releasing the enzyme in its reduced form. Structural perturbations necessary to favour the disulfide reshuffling are mediated by non-covalent interactions between the target protein and different domains of the catalyst (56,57). Cys_1 and Cys_2 in the active form of DsbA are linked with a disulfide bridge. The DsbA active site is an electron acceptor which, receving two electrons from the reduced substrate protein, catalyzes the disulfide

formation. The reduced DsbA is then re-oxidized by DsdB (58) (Fig. 2). DsdB is a transmembrane protein presenting two periplasmic cysteine pairs located in two different domains. These two cysteine pairs are the electron acceptor which can oxidize the catalytic site of DsbA restoring its activity (59). Mutants in the DsbB gene producing an inactive protein accumulate DsbA in the reduced form and are deficient in disulfide formation (60). DsbB can finally regain its active oxidized form thanks to a not well-characterized membrane electron transportation pathway (61). Recent in vivo experiments demonstrated that disulfide bond formation is strictly related

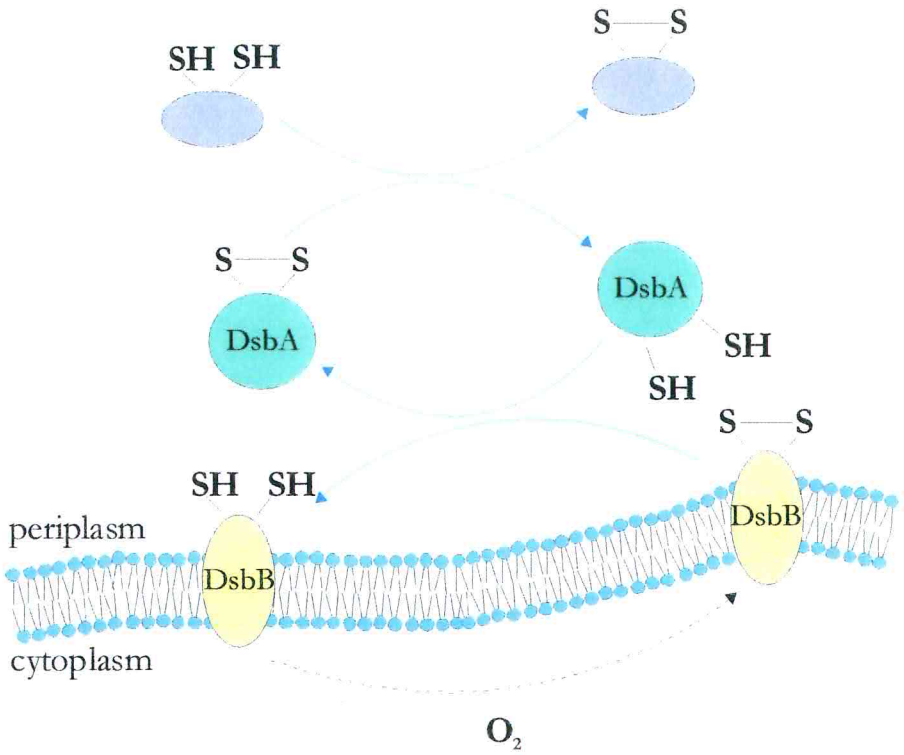


Figure 2: Schematic representation of the *In vivo* disulfide bond formation in *E. Coli*.

to the electron transport chain (60). The entire system relies on NADPH as a source of reducing equivalents, which is finally produced in the respiration process. We can then simplify the disulfide bond formation as an oxidation reaction between two cysteine residues and a molecule of oxygen.

$$2\,RSH + 1/2\,O_2 \longrightarrow RSSR + H_2O$$

Disulfide reshuffling in bacteria is catalyzed by a number of enzymes, among which a protein named DsbC is the most important (62). The role of these proteins is to correct the improper disulfide bonds produced during the process of oxidative folding. In *E. Coli* a complex cytoplasmic membrane protein called DsbD is responsible for regeneration of DsbC (53). DsbD is part of a poorly characterized electron transfer pathway which moves electrons from the cytoplasm to the periplasm.

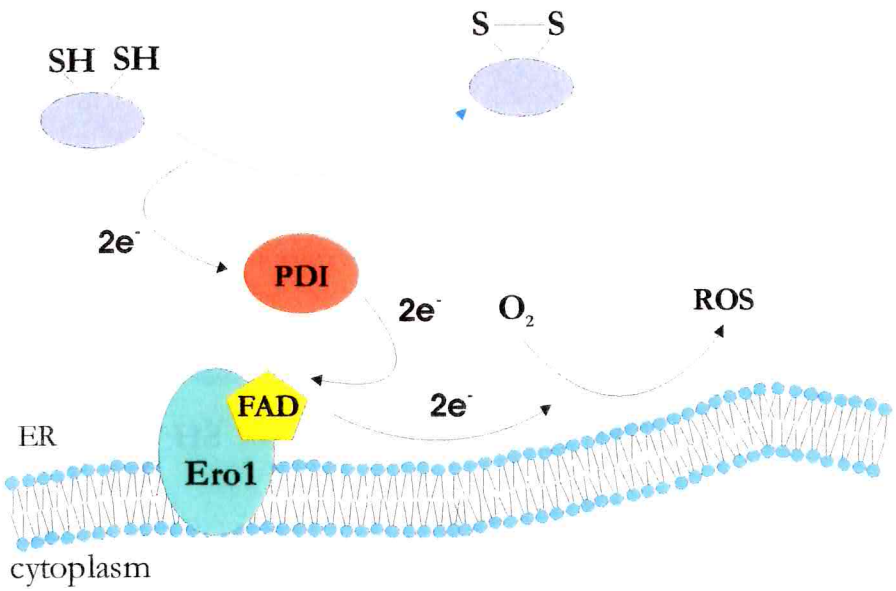


Figure 3: In eukaryotic organisms the oxidative folding takes place in the lumen of the ER. PDI and Ero1 are the most important catalyst of the disulfide formation and isomerization processes.

In eukaryotic organisms disulfide formation takes place in the lumen of the endoplasmic reticulum (ER). Protein disulfide isomerase (PDI) contains two thioredoxin domains and is considered the most important eukaryotic cathalyst both in disulfide formation and isomarization. In its oxidized form, PDI has the function of an oxidase for proteins that are emerging in the ER lumen while the reduced PDI released after disulfide formation can then catalyze the isomerization of incorrect disulfides. A conserved ER membrane associated protein (Ero1) have been demonstrated to be responsible for the direct oxidation of PDI playing a role analogous to that of the bacterial periplasmic protein DsbB (Fig. 3). While in bacteria the oxidative folding is coupled with the molecular oxygen through the respiratory chain, in eukaryotes Ero1 uses a flavin-dependent reaction to exchange electrons directly to the molecular oxygen. This process produces reactive

oxygen species (ROS) and is supported by an active transport system that imports FAD into the ER.

Many aspects of in vivo protein oxidative folding remain to be clarified. Very little is known about the exact mechanisms of electron transfer through membranes and many elements of the electron transfer pathways are still completely unknown. It is not clear yet how can PDI accomplish both oxidation and isomerization reactions, two processes that require a different oxidation state of the catalytic site.

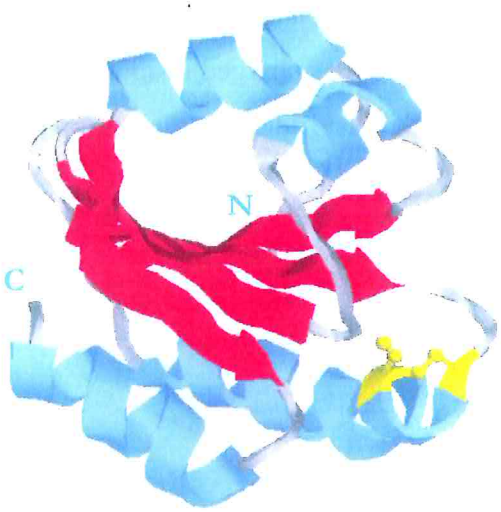


Figure 4: All proteins involved in disulfide-bond catalysis, in eukaryotes and prokaryotes, share a typical thioredoxin fold. The thioredoxin protein is a ubiquitous redox molecule. It consists in a single domain with a central five-stranded β-sheet (in red) flanked by four α-elices (in blue). The dithiol/disulfide active site is constituted by the sequential amino acids Cys_{32}-Gly_{33}-Pro_{34}-Cys_{35}. Cys_{32} and Cys_{35} are marked in yellow.

1.3.2 Disulfide bond formation in vitro

Oxidative folding experiments are performed *in vitro* in the presence of a redox couple that provides a correct equilibrium between reduced and oxidized forms. Generally the redox potential is generated by Cys/cistine or by reduced/oxidized glutathione (GSH/GSSG) couples. A thiolate anion of a free cysteine residue reacts with a disulfide bond of the redox couple forming a mixed-disulfide intermediate between the polypeptide and the redox reagent. The mixed-disulfide then reacts with another Cys to

generate a new disulfide bond (Fig. 5). The disulfide of the redox reagent can belong to the same protein, in this case it is an intramolecular reaction called disulfide reshuffling. The reactivity of the thiol group is a fundamental element in disulfide bond formation, and the pKa value of a thiol determines its degree of ionization and its intrinsic reactivity at any value of pH. The pKa of a typical cysteine sulfydryl group is about 8.6 and is strictly related with the chemical environment created, into the protein, in the close proximity of the sulfur atom. Since the disulfide formation/reshuffling process is promoted by the nucleophilic attack of a thiolate anion, the reaction rate will increase with the increasing pH and will be suppressed at acid pH. Non-polar environments and electrostatic interactions with nearby charged residues can heavily influence the reactivity of the thiol group and, thus, the disulfide bond formation (63).

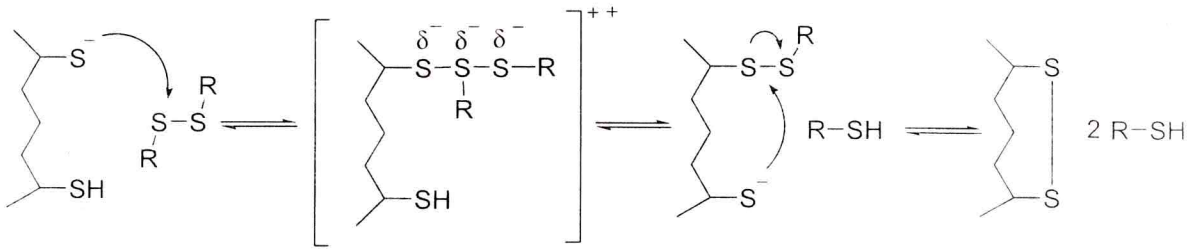


Figure 5: Disulfide bond formation and reshuffling occur always through a thiol/disulfide exchange reaction. The thiolate group is a strong nucleophile that can react with an existing disulfide bond. The reaction might involve a redox reagent present in solution (GSSG) or a pre-existing disulfide bond of the same protein promoting, in this second case, a disulfide reshuffling

The course of an oxidative folding reaction is conditioned by other two structural features proximity and accessibility (64). The accessibility of the disulfide bond by the redox reagent is a fundamental feature because the disulfide bond formation and reshuffling can occur thanks to a thiol/disulfide exchange reaction. A disulfide bond located at the protein surface can be formed rapidly, but thiol groups deeply buried in the structure have fewer possibilities to react. One the other hand even the opposite reaction of reduction is slow and disulfides buried in a folded protein are thus very stable. The proximity is determined by the tendency of the protein to bring the Cys in close approach with the right orientation during the folding pathway. The possibility for the sulfur atoms of two Cys residues to come in the correct position to react and form a disulfide bond depends on the energetic balance of the protein refolding process. Disulfide bonds, in fact,

constraint a intrinsically stable structure and their stabilization effect is, indeed, very complex depending on the structural environment. The structure perturbations necessary for the access of the redox agent impose a high free-energy barrier, and consequently the first step of the thiol/disulfide exchange reaction constitutes the rate limiting step. The formation of each disulfide bond can be hampered or favoured by the formation of the former one. The stabilization of a certain structure by a disulfide bond can allow or not the correct approach between other two thiol groups, can modify the chemical environment influencing their reactivity or impose a kinetic block burying them in the protein interior (49).

1.4 Protein folding

A protein, once synthesized, can correctly fulfill its biological function only if it first reaches the native conformation. A protein is able to autonomously search and find its unique native fold because the instructions for this task are included in its own amino acid sequence. Considering the different conformations that each amino acid residue can experience, even for a small protein the number of possible states is astronomically large. We could thus expect a protein to need an infinitive time to randomly find its unique correct fold. Surprisingly, the experimental observations demonstrate that proteins can find their way to the native state very quickly end effectively. Most of the single-domain proteins, in fact, can fold in vitro within a second or, very often, in a much shorter time (65).

All possible conformations form a potential energy surface where distinct states have different energies and are thus not equally probable. It is now accepted that the native state coincides with the condition of minimum energy and the protein can easily reach it following energetically convenient paths through a funnel-shaped energy landscape (66,67).

The classic idea of a specific and precise folding pathway has been replaced by the concept of the energy funnel where the folding is described as a statistical-based process. An unfolded protein sample is a heterogeneous population of molecules, each of them experiencing different conformations within all permitted states (68). According to the "molten globe" theory the polypeptide chain undergoes a rapid structural collapse led by the hydrophobic interactions. The resulting compact structure is stabilized by the formation of the hydrophobic cluster and by an array of specific and non-specific interactions between different parts of the peptide. This initial step occurs with a loss of potential energy due to the, on average, stabilizing character of the interactions. The "molten globe" corresponds with a local minimum of energy where the peptide is trapped while searching its native interactions. This is a bottleneck, slow step in folding process (49,69).

This stage is overcome, with an apparent two-state process, by the formation of a core structure that allows the system to precede downhill toward the native state (70).

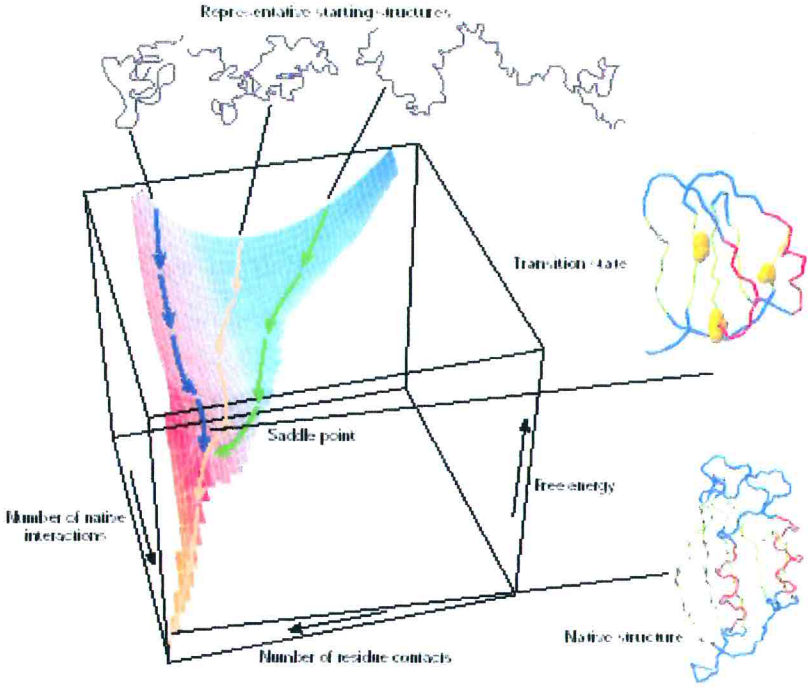


Figure 1: A schematic energy landscape for protein folding. The surface funnel is derived from a computer simulation of the folding of a highly simplified model of a small protein. The critical region on a funnel surface is the saddle point corresponding to the transition state, the barrier that all molecules must cross if they are to fold to the native state. Superimposed on this schematic surface are examples of structures corresponding to different stages of the folding process. The structure of the native state is shown at the bottom of the surface; at the top are indicated schematically some contributors to the distribution of unfolded species that represent the starting point for folding of individual molecules. Also indicated on the surface are highly simplified trajectories for the folding of individual molecules. (Adapted from:(69))

If the folding pathway cannot be foreseen for a certain molecule because it is a statistical process, the energy landscape of a particular protein is self-included information, determinate by its own amino acid sequence. Distinct molecules of the population will travel on the same surface following different trajectories since the starting conformations are different but will all finally arrive at the unique native state (Fig. 1). The uniqueness of the native folding is due to the fact that the interactions that stabilize the native state, at the same time strongly destabilize misfolded structures with the same

sequence. The energy balance that favors the native structure sums up all contributions of the free energy function, entropy and enthalpy. Non-polar groups strongly favor the folded state owing to the hydrophobic effect and van der Waals interactions between their side chains in the tightly packed core (71). The contribution of polar groups (polar and charged side chains and peptide groups) is much smaller and its stabilization effect is a balance of interactions inside the protein core and outside with the solvent. The stabilization of the native state due to the contribution of enthalpy is counterbalanced by the configurational entropy, which strongly favors the denatured state. The result of these two contrasting contributions is that the native state of a protein has, in physiological conditions, a free energy that is only slightly lower than that of the denatured state.

1.4.1 Cotranslational protein folding

An important difference between *in vivo* and *in vitro* folding is that *in vivo*, during translation the N terminus of the peptide is available for folding before the C terminus. In contrast *in vitro* refolding takes place under conditions where the entire polypeptide is involved, since the beginning, in the reaction. In principle, the N-terminus can fold as it emerges from the ribosome or as it is translocated into the ER. The successfulness of *in vitro* refolding experiments demonstrates that the cotranslational folding is not essential to obtain a correctly folded protein. *In vivo* folding process must be completed only once the protein is fully synthesized in order to establish multiple highly cooperative interactions essential to stabilize the tertiary structure. The cotranslational and cotranslocational folding involves other complex processes like the disulfide bond formation and the post-translational modifications. The whole process is assisted *in vivo* by enzymes and molecular chaperons, which are located in the lumen of the ER. Unlike *in vitro* refolding of small single domain proteins, the refolding of larger multidomain proteins *in vitro* is generally inefficient. Single domains can usually be successfully refolded individually in an *in vitro* process, whereas the same domain within the full-length protein is unable to fold properly. These observations suggest that the folding of an individual domain may be hindered by unfavourable interactions in the rest of the

protein. The difficulties in refolding large proteins raise the possibility that, the *in vivo* folding process of large multidomain protein is promoted by the cotranslational (or cotranslocational) formation of folded structures. This theory is known as the vectorial model. It describes the cotranslational folding as a systematic process that proceeds linearly from the N to the C terminus. Recent researches are shedding a new light on the cotranslational folding demonstrating that for the low-density lipoprotein receptor (LDL-R) the folding process does not occur in vectorial manner. LDL-R is a multidomain protein that reaches its native state passing through misfolded states containing non-native long-range interactions involving even incorrect disulfide pairings between distant cysteines. The progression of folding appears to depend on the balance of local and non-local interactions and to involve a rapid collapse to misfolded states that require re-organization to reach the native state. In the case of LDL-R and perhaps in other multidomain proteins, the vectorial model seems to be not the correct one. The protein does not reach its native state through the folding of smaller independent units but through a rearrangement of the full-length molecule. The capacity of the protein to fold according the correct frame therefore lies into the different modules. According to the vectorial theory the independent domains are guided toward the correct structure through a limited combination of local interactions. On the contrary, in the non-vectorial model, the modules can attain their correct structure into an array of different folding frames and log-range interactions. It is not clear if the native structures are selected according to their kinetic (rapidity of folding) or for their thermodynamic (minimum energy) features.

1.5 Tenascin-C and tenascin family

The tenascin family is composed of three large glycoproteins (~240 KDa) present in the extracellular matrix and known as Tenascin-C, -R and –X (41). These proteins are expressed during embryonic development in brain, cartilage, mesenchyme and are re-expressed in tumors, wound healing and generally, in adult tissues under active remodeling. Their localization pattern and the high level of conservation in all vertebrate species suggest that tenascins play an essential role in cell adhesion and motility (42). Tenascin-C, as the other members of the family, is found in tissues as a disulphide-linked hexamer called hexabrachion. The Tenascin monomers are modular proteins constituted of a series of repeated domains. All the members of the family show a characteristic pattern of four domain types. The 150 amino-terminal residues constitute the distinctive region that contains the oligomerization domain. Next to it a series of EGF-like domains each 31-34 amino acids long that can contain from 8 up to 18 modules. Remarkably, this region is encoded by a single exon. The EGF-like array is preceded by an incomplete EGF module. The next segment is made up by a series of fibronectin type III domains (FN-III) each of them about 91 amino acids long. A significant characteristic of tenascins is that splice variants exist for each of the three forms, and are characterized by the difference in the number of FN-III modules, suggesting that these domains have a functional role that can be regulated by splicing. In TN-C and TN-R, the alternatively spliced FN-III modules are inserted between domains 5 and 6.

The carboxy-terminal domain, which is highly conserved in TN-C, TN-R and TN-X, consists of 215 amino acids and is homolog to the C-terminal domain of β and γ-fibrinogen.

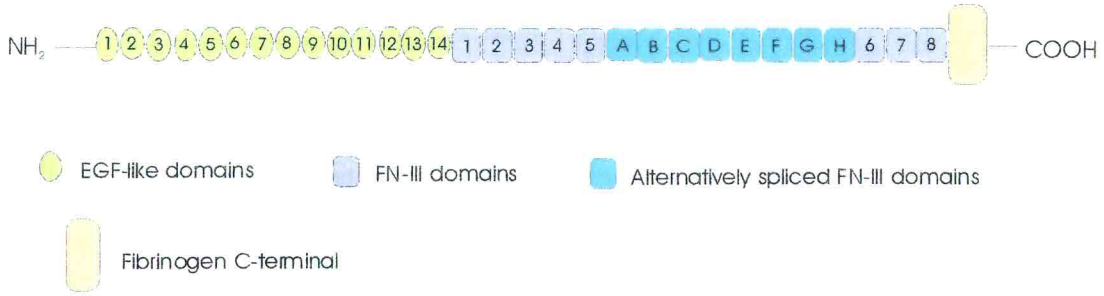


Figure 1: Tenascin C is a modular protein constituted by a series of repeated, independently folding domains. The number of FN-III domains is variable and distinguish different splicing isoforms. Alternatively spliced domains are inserted between modules 5 and 6.

One of the most important findings about this protein family is the re-expression of TN-C in several kinds of tumors. TN-C is largely expressed in malignant breast cancer (43), in a variety of benign tumors (44,45) and in fibrocystic disease, raising the hypothesis of a clinical relevance of this protein (46). Very recently, it was demonstrated that specific EGF-like modules of TN-C can elicit a mitogenic response through the EGFR signaling pathway (27). EGF-like domains 11/12/13 and 14 of TN-C have been demonstrated to act as "immobilized" ligands of the EGFR promoting mitogenic activity at a concentration range between 3 and 0.3 μM. The mitogenic stimulation has been demonstrated even in the absence of a soluble EGF factor. These data support the hypothesis of an intrinsic capacity of the EGF-like domains to act as low affinity ligands of the EGF-receptors.

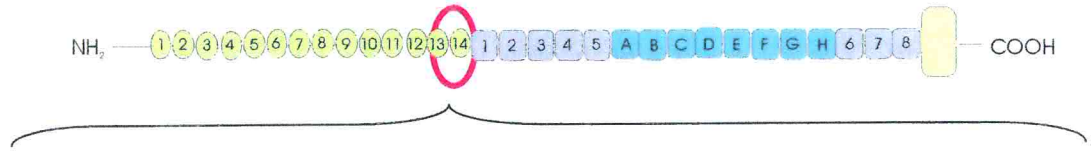
1.6 Aim of the work

Epidermal growth factor-like domains (EGF-like) constitute a heterogeneous family characterized by a highly conserved disulfide topology. Beside the invariable disulfide bond pattern, the length and the amino acid composition within the family is widely variable resulting in the peculiar capacity of EGF-like modules to adapt different sequences into the same structure. One of the most important structural features of EGF-like family is a two-stranded antiparallel β-sheet connected by two disulfide bounds (1-3 and 2-4) with the N-terminal part of the protein. The resulting structure is rather compact and stable with the cysteine residues involved in disulfide bond closely paired to one another. The EGF-like domains are present in many different proteins, many of which belong to the extracellular-matrix and contain several consecutive copies of EGF-like modules. Human Tenascin-C for example has an array of 14 EGF-like domains whose correct folding depends on the proper combination of all their 84 cysteines residues. All modules, in fact, have the typical EGF-like disulfide topology and structure, and all cysteines have to pair with their correct counterpart within each repeat.

```
187 EPECPGNCH.LRGRCIDGQCICDDGFTGEDCS      EGF-1
    QLACPSDCN.DQGKCVNGVCICFEGYAGADCS      EGF-2
    REICPVPCSEEHGTCVDGLCVCHDGFAGDDCN      EGF-3
    KPLCLNNCY.NRGRCVENECVCDEGFTGEDCS      EGF-4
    ELICPNDCF.DRGRCINGTCYCEEGFTGEDCG      EGF-5
    KPTCPHACH.TQGRCEEGQCVCDEGFAGLDCS      EGF-6
    EKRCPADCH.NRGRCVDGRCECDDGFTGADCG      EGF-7
    ELKCPNGCS.GHGRCVNGQCVCDEGYTGEDCS      EGF-8
    QLRCPNDCH.SRGRCVEGKCVCEQGFKGYDCS      EGF-9
    DMSCPNDCH.QHGRCVNGMCVCDDGYTGEDCR      EGF-10
    DRQCPRDCS.NRGLCVDGQCVCEDGFTGPDCA      EGF-11
    ELSCPNDCH.GQGRCVNGQCVCHEGFMGKDCK      EGF-12
    EQRCPSDCH.GQGRCVDGQCICHEGFTGLDCG      EGF-13
    QHSCPSDCN.NLGQCVSGRCICNEGYSGEDCS 621  EGF-14
       *   *       *    * *         *
```

Figure 1: Sequence alignment of the 14 EGF-like sequence domains of Tenascin-C

The aim of this work is to contribute to understand which are the determinants that drives such a multidomain protein in its correct folding frame. In order to simplify the system under study we synthesized a serie of 6 peptides of 33 amino acids containing 6 cysteine residues. Each peptide corresponds to a part of the tenascin sequence spanning from residue 560 to residue 622 that accomodates EGF-like modules 13 and 14. Starting with the EGF-like 14, the sequence of the peptides is obtained sliding the 33 amino acids reading frame by one cysteine at the time toward the N terminal of the protein. In this way all peptides contain 6 cysteines differently spaced on a native EGF-like amino acid sequence.




```
EQRCPSDCHGQGRCVDGQCICHEGFTGLDCGQHSCPSDCNNLGQCVSGRCICNEGYSGEDCSE
                                GQHSCPSDCNNLGQCVSGRCICNEGYSGEDCSE   f1
                     EGFTGLDCGQHSCPSDCNNLGQCVSGRCICNEG              f2
                ICHEGFTGLDCGQHSCPSDCNNLGQCVSGRCI                    f3
           DGQCICHEGFTGLDCGQHSCPSDCNNLGQCVSG                        f4
      HGQGRCVDGQCICHEGFTGLDCGQHSCPSDCNN                             f5
 SDCHGQGRCVDGQCICHEGFTGLDCGQHSCPSD                                  f6
```

Figure 2: Sequence of the peptide under study and their position into the EGF-13 and EGF-14 modules

Our approach intends to investigate whether the native EGF structure is favored by the kinetics of the folding process or rather if are the structural elements (cysteine spacing) that impose a unique possible fold. The model simulates a multidomain system and permits to compare folding process of a native in frame EGF with that of the frame shifted peptides, forced to an out-of-frame misfolding.

2 Materials and Methods

2.1 Synthesis of EGF-14

The difficulty of the synthesis was initially assessed with the Peptide Companion software and through consultation of the bibliography regarding the synthesis of similar peptides. It was thus decided to carry out the synthesis of EGF 14 (Swiss-Prot: TENA_HUMAN) automatically with the PS3 Protein Technology synthesizer available in our laboratory. The 33 amino acids peptide was build on 350 mg of TentaGel S PHB-Glu(t-Bu)Fmoc resin with a substitution of 0.2 mmol/g to obtain a 0.07 mmol synthesis scale. The standard Fmoc-based strategy was employed. The incoming Fmoc-protected amino acids were used in 4 molar excess with respect to the original resin substitution. The final concentration of the protected amino acid in the reaction mixture was 0.3 M. The uronium salt activator TBTU was chosen as coupling reagent and used in 1:1 ratio with the protected amino acid. The neutralization of the reaction mixture was achieved with N,N-Diisopropylethylamine (DIEA) used in 8 molar excess with respect to the growing peptide. All coupling reactions took place in DMF and were carried out at room temperature for 45 minutes until the residue Cys16. Reaction times were then prolonged to 1.5 h for the remaining amino acids. This strategy was followed to compensate the increasing difficulty of amino acid incorporation consequent to the aggregation tendency of larger peptides.

A different procedure was adopted for the introduction of Cys residues in order to avoid racemisation. N-α-Fmoc-S-trityl-L-cysteine pentafluorophenyl ester [Fmoc-Cys(Trt)-OPfp] building blocks were manually added in DMF/DCM (3:1 v/v) solution at Cys positions. Coupling reactions with Fmoc-Cys(Trt)-OPfp were carried out for two hours at room temperature to compensate for the low reactivity typical of preformed active esters. The Fmoc cleavage was achieved after each successful coupling with a double reaction of 5 minutes each. For the deprotection reaction a 20% piperidine, 0.1M HoBt solution was used.

Once the synthesis was complete, the protected peptide-resin was washed with DCM, dried and stored at 4 °C. The cleavage reaction proceeded for 2 h at room temperature in 90% (v/v) TFA, 5% (v/v) EDT, 2.5% (v/v) TIS, 2.5% (v/v) water and 0.5 M phenol. The reaction mixture was then filtered in order to separate the insoluble resin from the solution containing the peptide and the scavengers. The solution was purged with a nitrogen stream, evaporated in vacuo and the pellet dissolved in water. The scavengers were removed from the aqueous solution by mean of five extractions with 8 volumes of diethyl ether. The peptide solution was finally freeze-dried and stored at 4°C.

2.2 Manual synthesis of frame-shifted peptides

Whereas the synthesis of an egf-like peptide is a well-established procedure, the preparation of the frame-shifted products was actually new and difficulties could only partially be assessed referring by similarity to the EGF-like peptides. We thus decided to synthesize simultaneously the 5 frame-shifted peptides using a manual procedure. With this procedure the formation of deletion products originated by unexpected difficult couplings can be minimized. In Manual synthesis the completeness of each synthesis step can be verified before the incorporation of the next amino acid. Moreover, in a manually performed synthesis, different peptides can be prepared at the same time significantly reducing the time required. The peptides were built on TentaGel-S-PHB solid supports derivatized with the proper Fmoc protected amino acids. The coupling reactions were performed with a 4 molar excess of Fmoc protected amino acids, with respect to the growing peptide, and the same excess of activator reagent (TBTU). The coupling reaction took place in DMF in the presence of 8 equivalents of DIEA for 1 h at room temperature. The completion of each coupling reaction was proved with the Kaiser assay. This is a sensitive colorimetric test that reveals the presence of free amines in the growing peptide if an incomplete acylation reaction occurred (72). A second coupling reaction was attempted, if revealed necessary by a positive result of the Kaiser's test. Stronger coupling reagents like PyBop or HATU were employed in the second reaction. Fmoc protecting groups were removed by a 20% (v/v) piperidine solution in DMF. HOBt was

added in 0.1 M concentration to the deprotection solution to minimize the occurrence of aspartimide formation.

Even for these peptides the incorporation of Cys residues was achieved employing the activated ester of the amino acid. Fmoc-Cys(Trt)-OPfp was used in the first acylation reaction, which was carried on for 2 h at room temperature. The double coupling procedure was systematically adopted for Cys residues. In the second coupling reaction Fmoc-Cys(Trt)-OH building block was used with TBTU as coupling reagent and 2,6-dimethylpyridine as weak base to minimize racemization.

After the first Cys residue of each peptide a capping reaction was performed. This step was introduced to block the growth of deletion products on unfunctionalized binding sites on resin support. Capping was repeated twice at room temperature with a 0.5M Ac_2O, 0.125M DIEA and 0.2% (w/w) HOBt solution in NMP.

In peptide f3, two extra amino acids have been added (Ser at the carboxyl and Ala at the amino terminal) to prevent a cysteine residue to be located at the N/C terminus of the peptide. Such a position might affect the possibility to interact with other Cys residues to eventually form disulfide bonds. The proximity of the thiol group with the terminal charge can modify its reactivity. At the same time the N/C charge can influence, by electrostatic attraction or repulsion, the approaching with other Cys residues. Finally, a terminal Cys is much less sterically hindered with respect to an internal one; the major accessibility can favour the disulfide bond formation. Alanine and Serine were chosen because supposed to cause limited perturbations of the physico-chemical characteristics of the peptide. Once the synthesis was completed the product was cleaved/deprotected with TFA following a standard protocol, extracted six times with diethyl ether and finally freeze-dried. Detailed records of the syntheses are reported in Appendix A.

2.3 Peptide purification

The crude egf-14 and f5 peptides were purified by RP-HPLC on a Gilson chromatographic apparatus using a PrePak Cartridge 25 x 100 mm C-18 (Agilent) cast on a PrepLC Universal Base apparatus (Waters) with a linear gradient of triethylammonium

acetate buffer (25 mM, pH 7) and triethylammonium acetate buffer (25 mM, pH 7) in water/MeCN 1/9 (v/v). The peptides were further purified and desalted. This was achieved through a second RP-HPLC on the same Gilson chromatographic apparatus with a Zorbax 300SB-C18 9.4 x 250 mm column (Agilent). Samples were eluted with a linear gradient of water/TFA 0.1 % (v/v) (buffer A) and MeCN/TFA 0.1 % (v/v) (buffer B).

Preparative HPLC of frame-shifted peptides f2, f3, f4 and f6 were carried out with the same chromatographic apparatus (Gilson) with a PrePak Cartridge 25 x 100 mm (Agilent) and Zorbax 300SB-C18 9.4 x 250 mm (Agilent) columns. For these peptides a linear gradient of buffer A and buffer B was used.

2.4 Folding and purification of EGF-14

The crude EGF-14 peptide prepared as previously described underwent a large scale folding process. The pure peptide was desalted by a RP-HPLC (see above) with a H_2O/MeOH/0.1% TFA buffer system and then freeze-dried in the presence of NaOH in order to remove traces of TFA. Lyophilised EGF-14 was dissolved in acidic water solution (0.01% TFA) to prevent disulfide formation and then diluted 10 times in the refolding buffer (0.1 M ammonium acetate, 2 mM EDTA, Cys/cystine 20:1 (w/w), pH 8.5) previously purged with a nitrogen stream. The final peptide concentration was 50 mg/L while the molar ratio between peptide cysteines and the cystine in the redox couple was 1:10. The folding reaction was carried out over night at room temperature even if HPLC-MS controls demonstrate that the process was complete after 2 hours. The folded EGF-14 was finally purified by preparative RP-HPLC with a C-18 PrePak Cartridge 25 x 100 mm (Agilent) on a Gilson chromatographic apparatus. Folded EGF-14 was eluted with a linear gradient of buffer A and buffer B and finally lyophilised.

2.5 Time course folding experiments

After purification by HPLC, the reduced and lyophilized peptides (egf-14 and f2-f6) were dissolved in an acidic water solution (TFA 0.01% (v/v)) in order to prevent, with the low pH, an uncontrolled rearrangement of the disulfide bonds. The time-course folding experiment is started diluting the reduced peptides 10 times in the refolding buffer (0.1 M ammonium acetate, 2 mM EDTA, Cys/cystine 20:1 (w/w), pH 8.5). In order to eliminate the oxygen, which can alter the redox potential of the solution, the refolding buffer is previously flushed for 10 minutes with a nitrogen stream. The molar ratio between the peptide cysteines and the cystine of the redox couple in the refolding buffer was 1:10. Comparable amounts of each peptide, estimated by HPLC-UV detection at 214 nm, were used in the time course refolding experiments that took place at room temperature in a final volume of 5 mL. Aliquots of the reaction mixtures were quenched at selected times (2.5, 5, 10, 15, 20, 30, 40, 60, 90, 120 min, 4 h and 24 h) by acidification with TFA (2% (v/v) final concentration) and immediately stored at –80 °C. The folding progression was then monitored by LC-MS (see Mass spectrometry section).

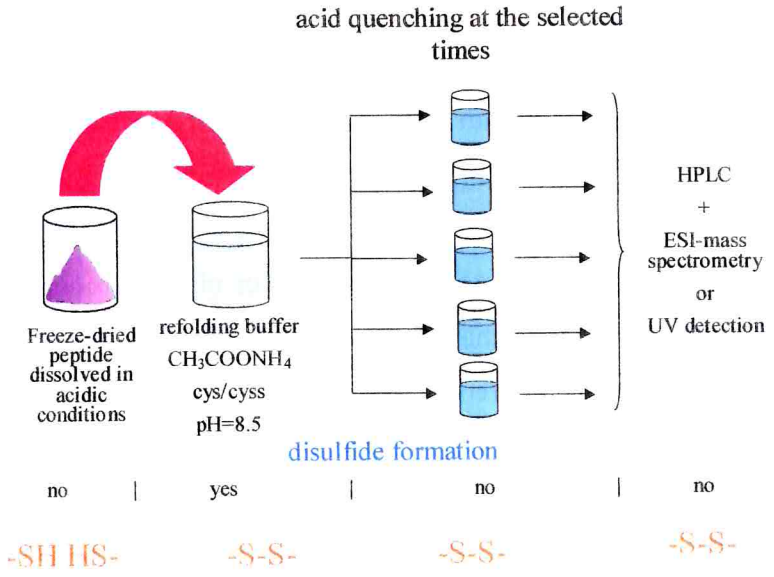


Figure 2: Schematic representation of the time-course refolding experiment. The oxidative state of the cysteine residues is shown in the lower panel

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 x 150 mm, 3.5μm) column by means of the same buffer system (buffers A and B).

2.6 Disulfide bond topology

The assignment of the disulfide bond pattern was achieved by the analysis of the fragments produced by digestion with properly chosen endoproteinases. The initial hypothesis for the egf-14 peptide was a set of disulfide bridges typical of egf-like domains. The digestion strategy was then set up to verify or discharge the starting assumption. Egf-14 was first treated with thermolysin. 40 μg of the purified peptide were dissolved in 250 μl of 10 mM sodium borate buffer (pH 6.0) and incubated at 37°C for 2, 5 and 12 hours. The reactions were carried out with two different enzyme/substrate ratios (1:3 and 1:10 w/w). In order to unequivocally assign the disulfide topology, a second digestion was necessary. A part of the digest (enzyme/substrate 1:3, 12h) was then adjusted to pH 6 with sodium phosphate buffer and AspN endoproteinase was added to perform the second cleavage, and the reaction mixture was incubated for 12 h at 37 °C. In both sets of digestions, reactions in absence of substrate and in absence of enzyme were contemporarily performed as controls. Digestion mixture were analysed by RP-LC-MS using a Zorbax 300SB C18 column (1.0 x 150 mm, 3.5 μm, Agilent). Different linear gradients and MS parameters were employed in order to optimize the methodology according to the specific conditions. Prediction of digestion products was obtained using ProMAc (Applied Biosystem) and Sherpa light 4 (73) Macintosh based software.

2.7 Circular dichroism spectroscopy

The secondary structure of folded EGF-14 and of the main folding products of f5 and f6 was investigated by Circular Dichroism spectroscopy. Samples were prepared dissolving the lyophilized peptides in 500 μL of milly Q water. 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, with a scan speed of 10 nm/sec and the baseline subtracted from the raw spectra. The molar ellipticity was calculated from the CD signal intensity (mdeg) divided by c x l, where c is the concentration (M) and l the path length (cm). Peptide concentration was determined by amino acid analysis and UV absorption at 215 nm.

2.8 Amino acid analysis

The concentration of the samples analysed by CD were estimated with the amino acid analysis methodology. For this procedure a PICO-TAG Amino Acid Analysis System (Waters) was employed following the suggested standard protocol. Lyophilised peptides were hydrolysed under HCl saturated atmosphere for 1 hour at 150 °C in presence of crystals of phenol. Samples were dissolved in re-drying solution (ethanol/water/triethylamine, 2/2/1 (v/v)) and lyophilised again. This procedure was repeated twice in order to reach the basic pH required by the derivatization reaction. Derivatization solution (PITC/EtOH/TEA/water in ratio 1/7/1/1 by volume) was added in the vials and the reaction let to proceed at room temperature for 20 min. Samples are then freeze-dried and stored at -80 °C. RP-HPLC analyses were carried on with a gradient of acetonitrile in Na-acetate buffer (pH 6.4) with a PICO-TAG column thermostated at 46

°C operating with a Gilson 306 pump system. UV detection was set at 254 nm. The amount of the amino acids was calculated referring to a calibration curve obtained with a commercial standard mixture of 19 amino acids (Fluka). The measurement of the peak areas was performed with the Gilson UniPoint software.

3 Results

3.1 Peptide synthesis

The purity of the crude products was estimated by RP-HPLC with UV detection at 214 nm. LC-MS analysis confirmed, for all peptides, the correspondence between the molecular weight assigned to the main chromatographic peaks and that of the expected products (Table1). Side products originated by piperidine transamination are the most common contaminants heavily affecting particularly the synthesis of peptides f5 and f6. Other side products are mainly due to deleted peptides originated by not-quantitative incorporation reactions. Among them, des-Cys peptides were particularly frequent. Products of Cys racemisation were not observed.

peptide	method	yield (%)	purity (%)	final yield (%)	expected mass (Da)	observed mass (Da)
egf-14	a	85.5	66.4	56.7	3451.2	3452.0
f2	m	84.6	46.5	39.3	3483.4	3484.3
f3	m	86.1	44.1	38.0	3398.3	3398.0
f4	m	86.5	58.6	50.7	3327.3	3328.0
f5	m	83.4	48.3	40.3	3477.3	3478.7
f6	m	83.1	25.0	20.7	3451.3	3451.5

Table 1: Peptide synthesis. Synthetic method (a, automatic; m, manual), yield (%) of the crude deprotected peptide as estimated by weight; purity (%) of the crude deprotected peptide as estimated by HPLC; final yield (%); expected and observed average molecular mass (Da) of the reduced, purified peptide.

The employment of proper separation strategies (see materials and methods) permitted the recovery of the peptides with a purity ranging from 95 to more then 99% (Table 2).

Peptides	Final Purity
EGFf1	>99%
f2	99%
f3	99%
f4	99%
f5	99%
f6	>95%

Table 2: Purity of the peptides after RP-HPLC purification.

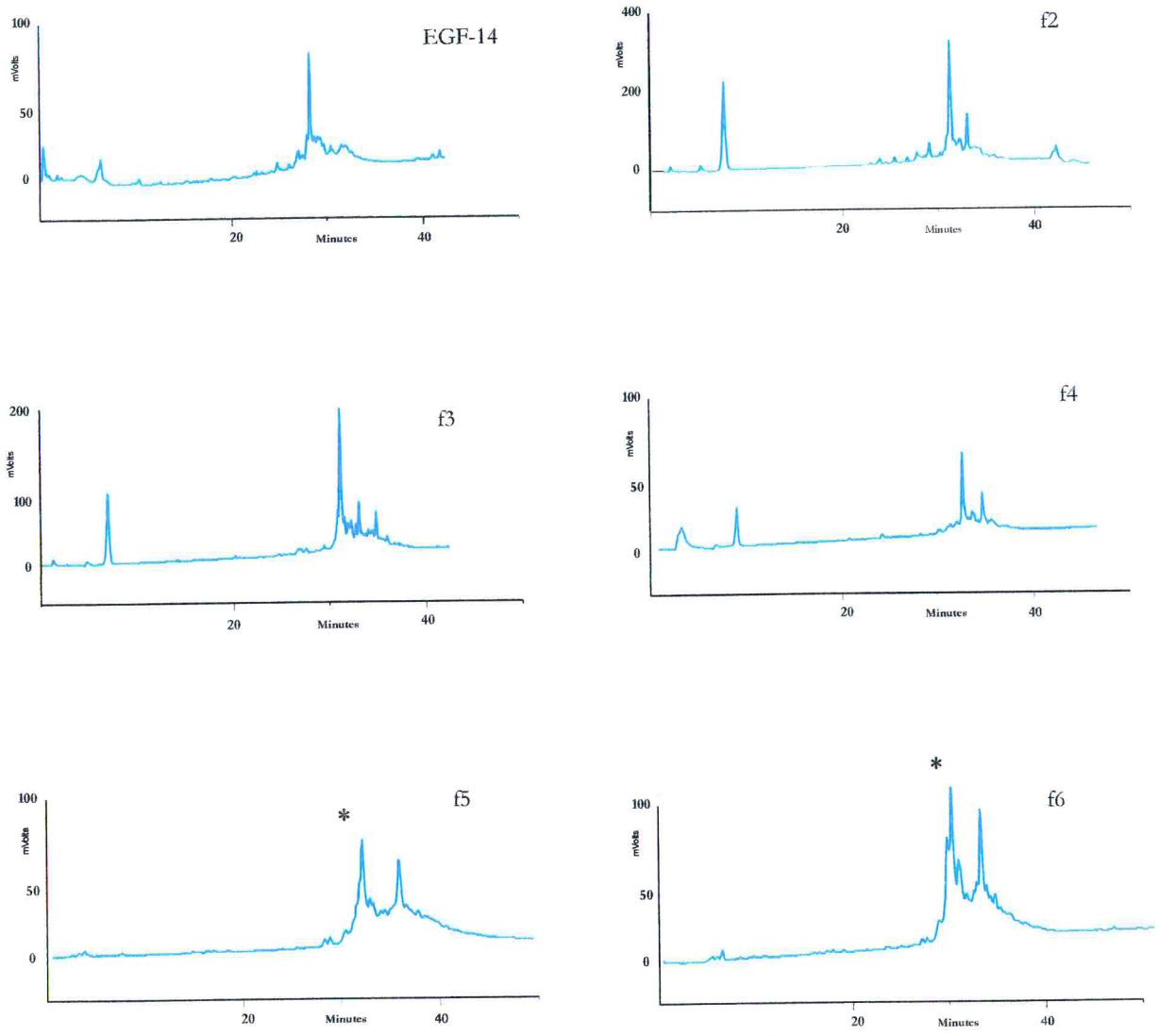


Figure 1: RP-HPLC profiles of crude peptides accomplished by UV detection at 214 nm. The stars indicate the peaks of peptides f5 and f6

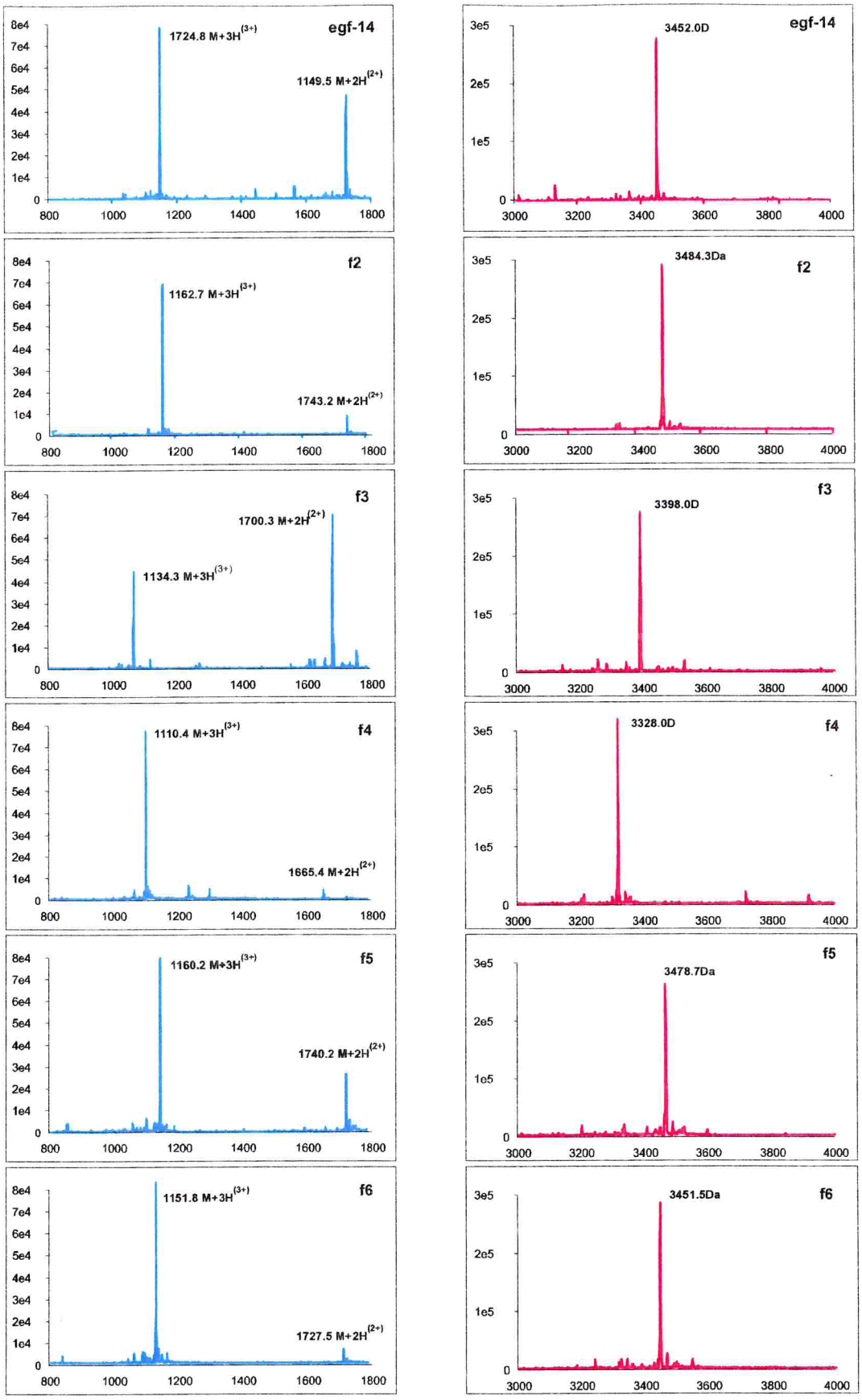


Figure 2: ESI-MS spectra of the peptides achieved in positive ion mode (left). On the right are shown the corresponding deconvoluted spectra

3.2 Oxidative folding

The purified, lyophilized peptides were refolded in the presence of the Cys/cystine redox couple as previously described. The time course of the refolding kinetics was followed by LC-MS (Fig. 3 and Fig. 4).

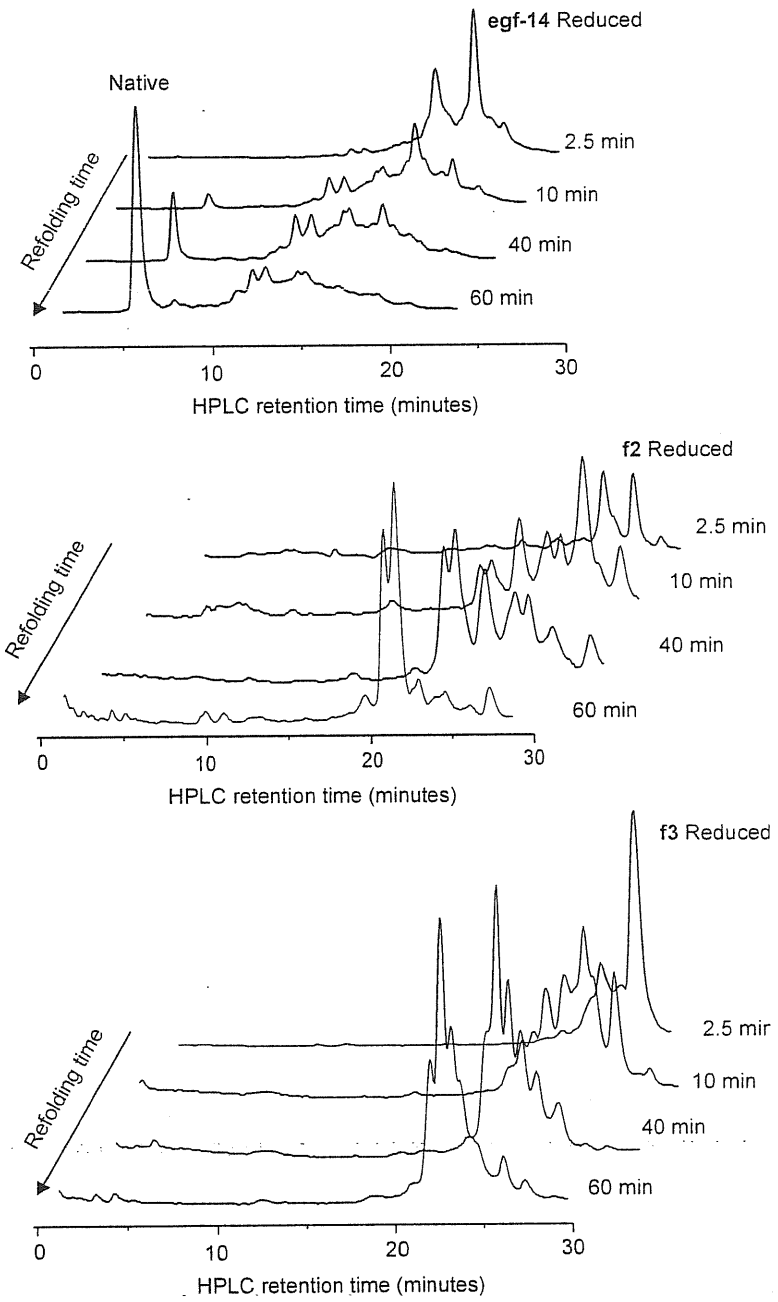


Figure 3: Oxidative folding. HPLC profiles of oxidative folding reactions, of peptides EGF14, f2 and f3 as detected by UV at 214 nm, of the different peptides at selected refolding times. The peak corresponding to the initial, fully reduced form is marked by the name of the peptide.

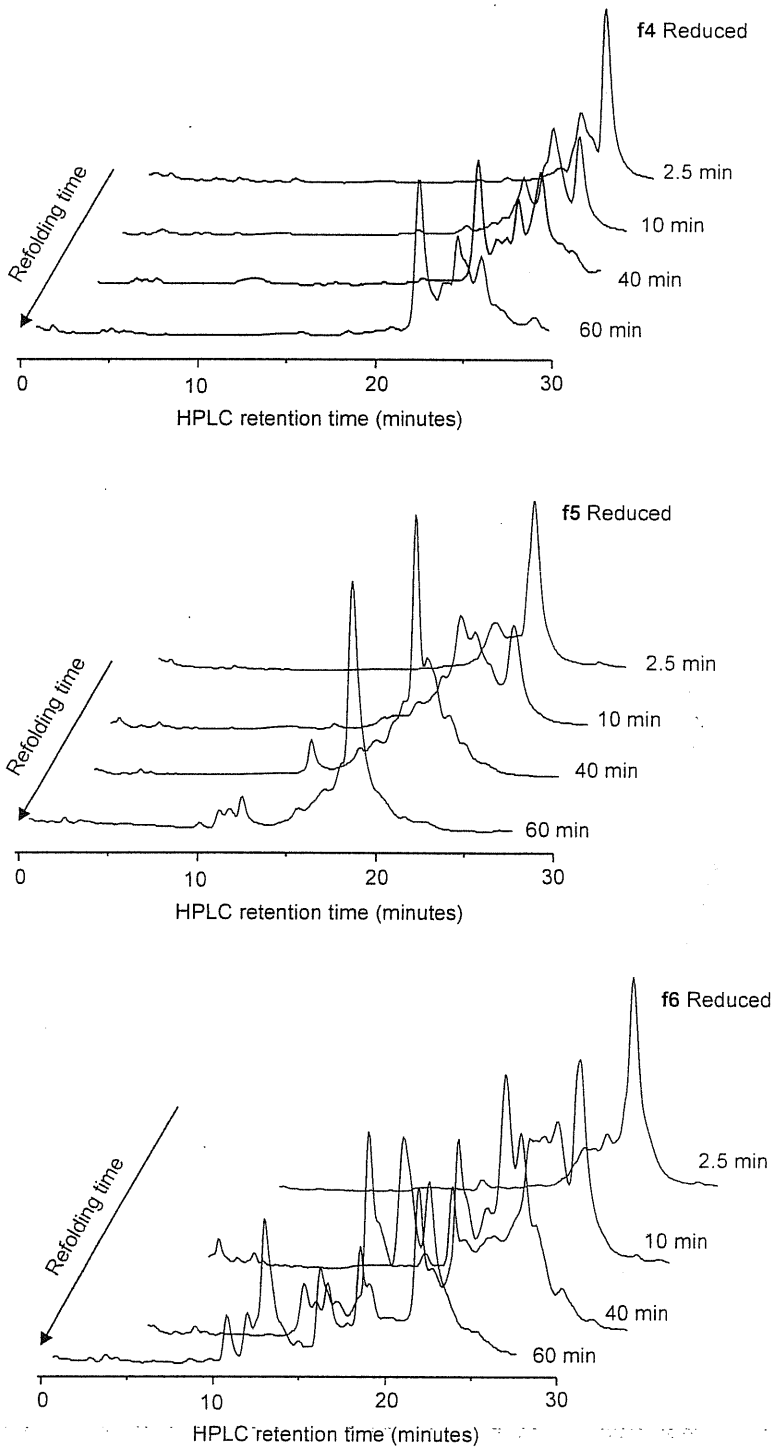


Figure 4: Oxidative folding. HPLC profiles of oxidative folding reactions, of peptides f4, f5 and f6 as detected by UV at 214 nm, of the different peptides at selected refolding times. The peak corresponding to the initial, fully reduced form is marked by the name of the peptide.

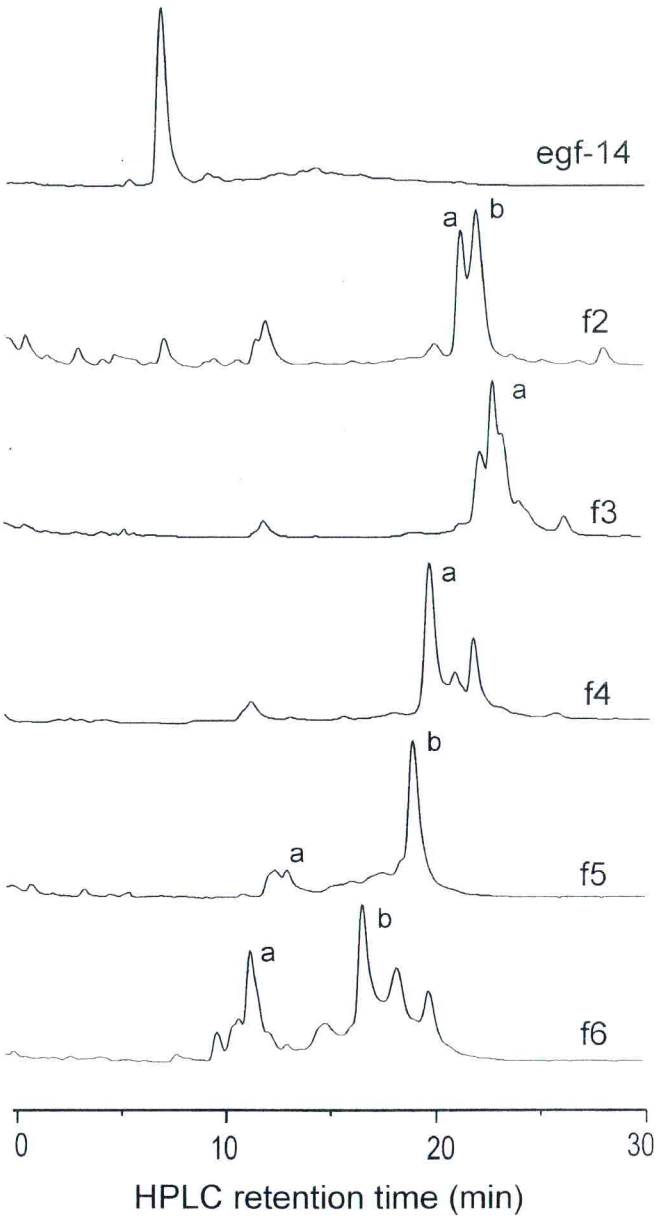


Figure 5: Equilibrium mixtures. HPLC profiles of the oxidative folding reactions after 24h. Detection was performed at UV at 214 nm.

LC-MS was used to monitor the formation of disulfide bonds from the loss of 2 amu in molecular mass for each disulfide formed, while HPLC was used to measure retention times and quantify the decrease of the starting product by UV detection at 214 nm. Experiments showed that the reduced peptides convert rapidly in a mixture of 1 and 2-

disulfide products, which undergo a slower oxidation and reshuffling to give several 3-disulfide isomers in frame-shifted peptides, while egf-14 gives a unique product.

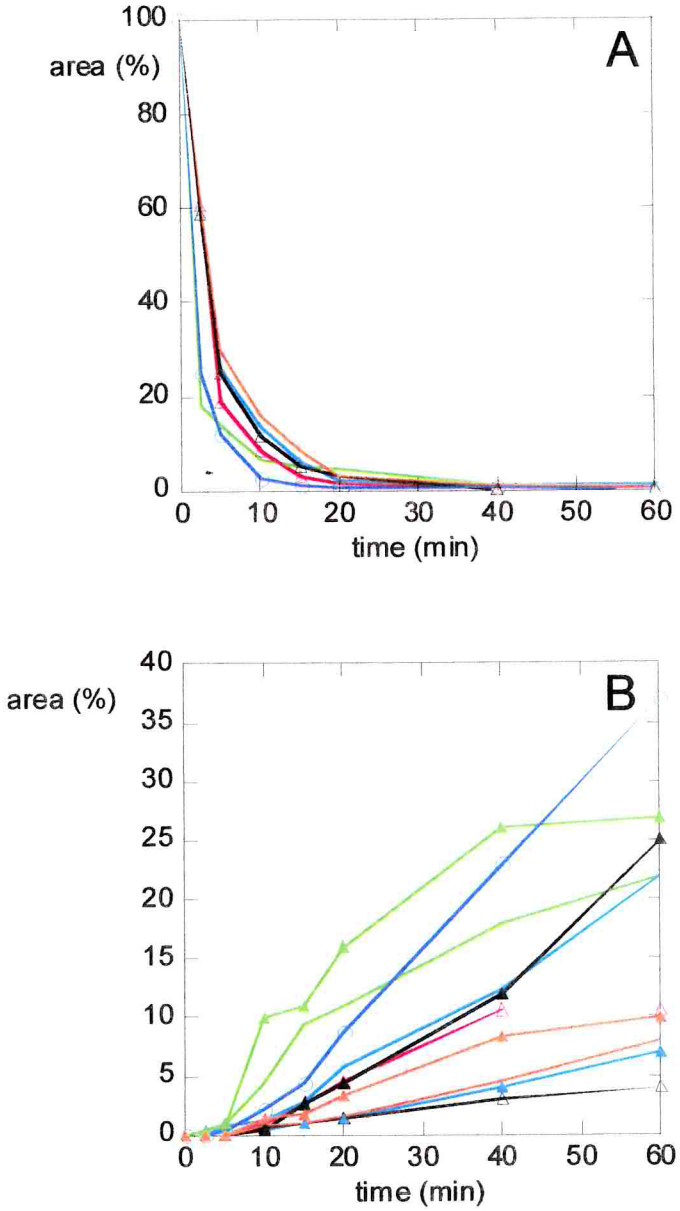


Figure 6: Oxidative folding kinetics. Panel A, disappearance of the starting product (%, area of the initial reduced form with respect to the total integrated area) for egf-14 (blue), f2 (green), f3 (red), f4 (light blue), f5 (black), f6 (orange). Panel B, formation of three-disulfide species (%, area of the 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) originating from the same peptide are shown as empty and filled triangles, respectively. Oxidative folding kinetics were followed by HPLC and UV detection at 214 nm

Under the described refolding conditions, efg-14 was rapidly oxidized and in two hours almost completely transformed into a unique 3-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 (Fig. 3 and Fig. 4).

The equilibrium pattern was reached for all peptides within 24 h (Fig. 5) and after this time changes in the relative abundance of the species or formation of new products were not observed (Data not shown). The LC-MS analysis confirmed that all products in the final mixtures are 3-disulfide isomers. The quantitative analysis of the HPLC profiles showed that the rates of disappearance of the reduced forms are similar but not identical (Fig. 6, Panel A). A three-parameter exponential fit of experimental data ($R > 0.99$) 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 that is smaller then that obtained for egf-14. A noteworthy difference in the rate of formation of 3-disulfide peptides was also observed (Fig. 6, Panel B).

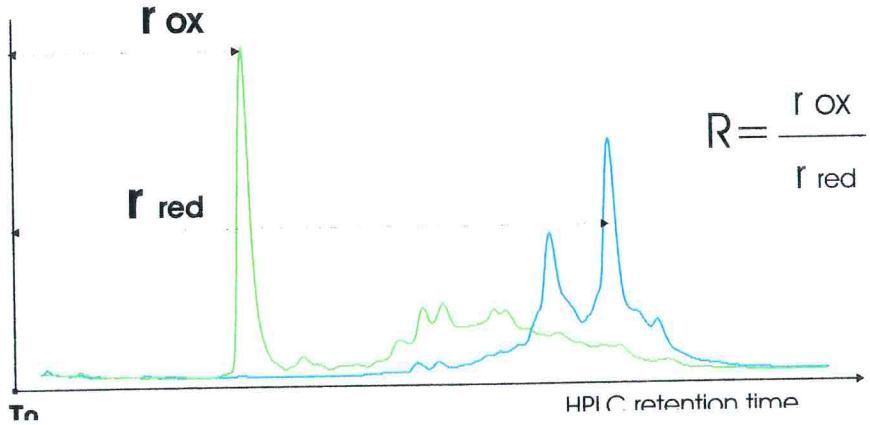


Figure 7: We define R as the ratio between the retention time of the oxidized species (r_{ox}) and that of the reduced one (r_{red}). Chromatograms relative to EGF14 are shown in the picture.

Egf-14 reached its fully oxidized form faster than the other peptides as demonstrated by LC-MS. A further difference between egf-14 and the frame-shifted peptides is represented by the change in the HPLC retention time (RT) going from the reduced to the oxidized species. The final product of egf-14 oxidative folding has a RT that is considerably shorter with respect to the reduced species (reduced form, 21.6 min; oxidized form, 7.7 min) (Fig. 7), while for frame-shifted peptides most products show RT values only slightly smaller than that of the corresponding reduced peptide. Only in the

case of f5 and f6, the RT of one of the final products is significantly reduced compared with the starting product. To quantitatively compare the behaviour of the different peptides, the chromatographic parameter R, defined as the ratio between the retention time of the oxidized product (r_{ox}) and the retention time of the reduced peptide (r_{red}) was chosen. As shown in figure 8, egf-14 displays the lowest R-value (Fig. 8).

peptide	RT_{red} (min)	RT_{ox} (min)	ΔRT (min)	R
egf-14	21.6	7.7	13.9	0.36
f2	28.7	f2a 21.6 f2b 22.3	7.1 6.4	0.75 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 f5b 18.2	10.5 4.6	0.54 0.80
f6	23.2	f6a 12.6 f6b 18.0	10.6 5.2	0.54 0.78

B

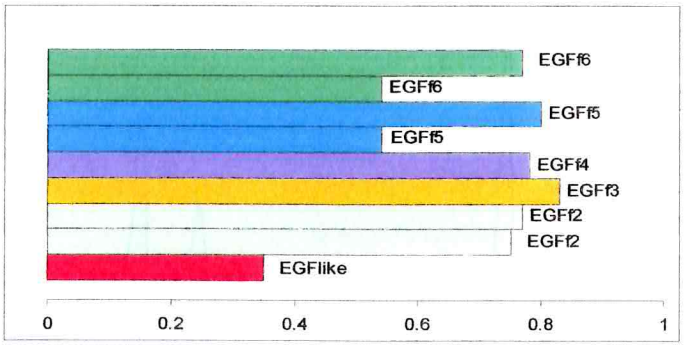


Figure 8: Values of the R parameter reported for the most represented reduced species of the different peptides. Table A reports the HPLC retention times. Retention times of the reduced (RT_{red}, min) peptides and of the main three-disulfide species (RT_{ox}, min); difference in retention times of the reduced and oxidized forms (ΔRT, min) and selectivity parameter (Δ defined as RT_{ox}/RT_{red}) for three-disulfide species. Panel B shows R values expressed in diagram.

3.3 Disulfide topology assignment

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 Cys22 and Cys31. The 2097 Da peptide, on the other hand, could not give an unequivocal answer about the two remaining bridges, which could be either Cys4-Cysl9/Cys8-Cysl4 or Cys8-Cysl9/Cys4-Cysl4. The 2097 Da peptide was therefore treated over night with AspN endopeptidase.

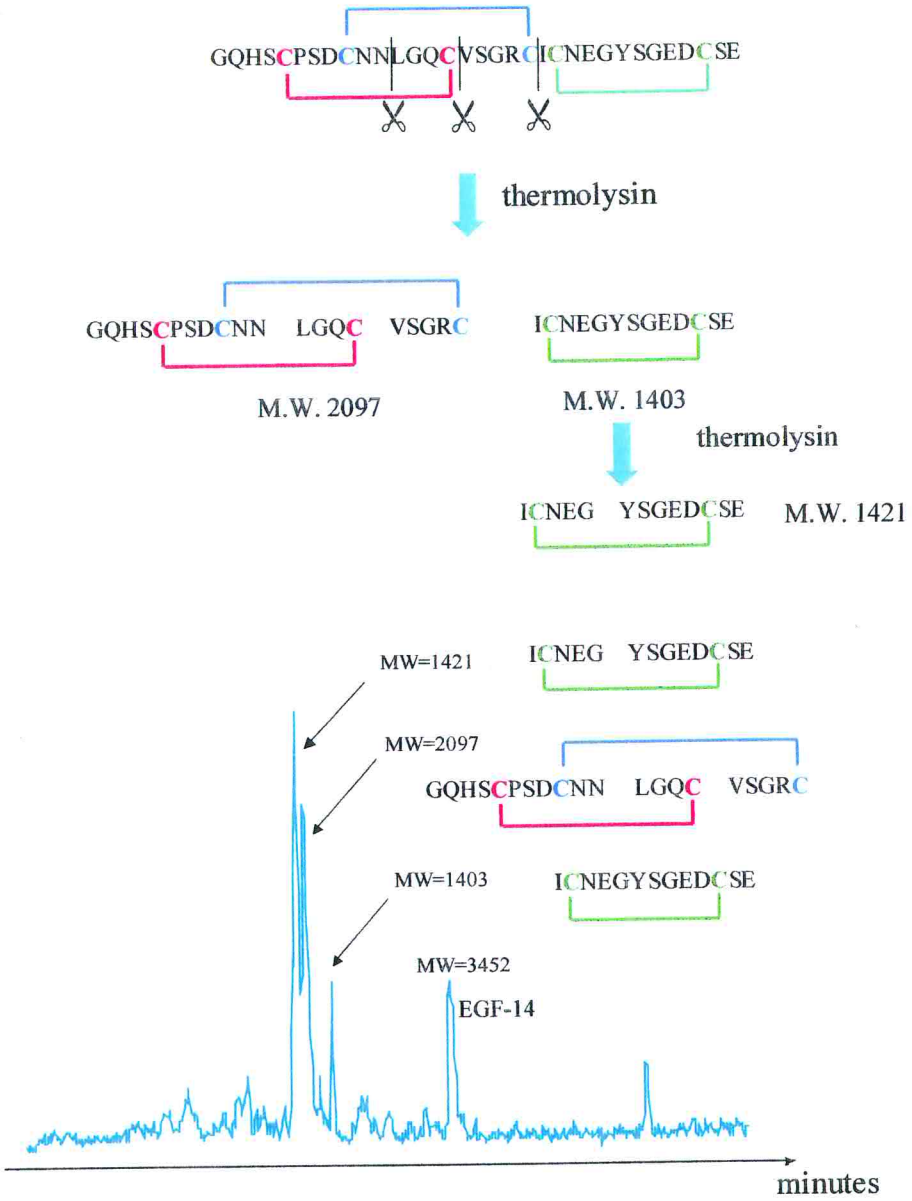


Figure 9: Peptide mapping strategy to determine the disulfide topology of egf-14. The thermolysin digestion confirms the 5-6 disulfide bridge but can not unambiguously define the topology of the first two bridges (Panel A). Proteolytic digestion is monitored by LC-MS (Panel B).

The reaction gave two fragments of 810 and 985.1 Da. This result unambiguously defines the Cys4-Cysl9/Cys8-Cysl4 combination (Fig. 10). The experiment thus confirms that the egf-14 from human tenascin has a disulfide topology typical of EGF domains (1,3-2,4-5,6).

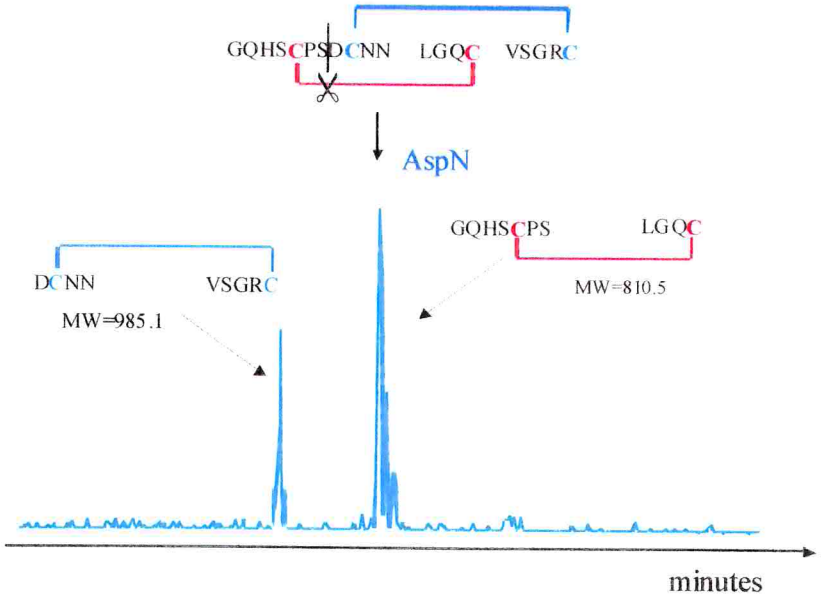


Figure 10: The 2097 MW fragment obtained by thermolysin digestion is treated with endoproteinase AspN and the digestion reaction is monitored by LC-MS. The reaction gave two fragments of 810.5 and 985.1 Da. This result unambiguously defines the Cys4-Cysl9/Cys8-Cysl4 combination. The experiment thus confirms that the egf-14 from human tenascin has a disulfide topology typical of EGF domains (1,3-2,4-5,6).

3.4 Circular dichroic spectra

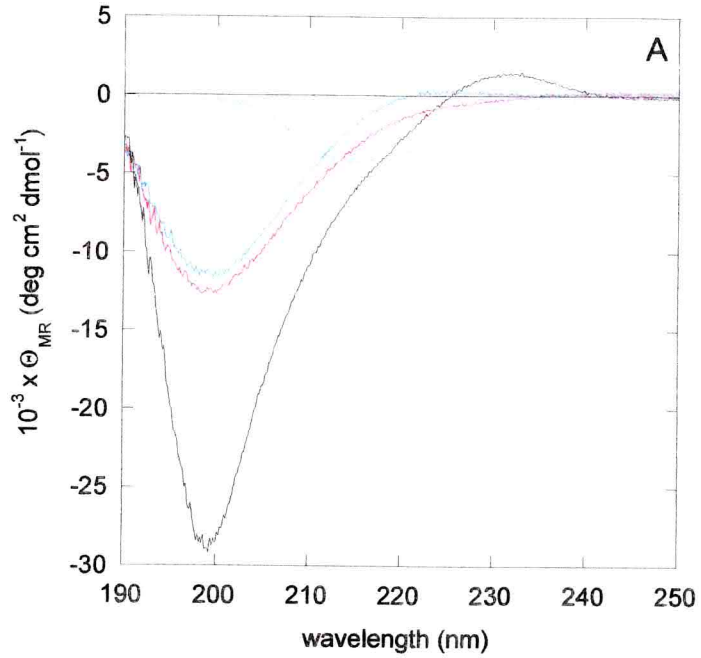
The CD spectrum of egf-14 is dominated by a negative band in the far-UV region (Fig. 11). This band has its minimum at 200 nm, a shoulder at 215 nm and is going to zero at ~190 nm. Two additional much weaker positive bands can be observed in the far-UV at ~235 nm and in the near-UV at 270 nm (Fig. 11B). The CD spectra of f5b and f6b (Fig11. A) 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 ~270 nm, and the intensity of this band is roughly four times weaker then that of egf-14. The CD spectra of f5a and f6a could be recorded only in the far-UV region. 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. 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.

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 (SELCON3, CONTINLL, CDSSTR, K2D). 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.



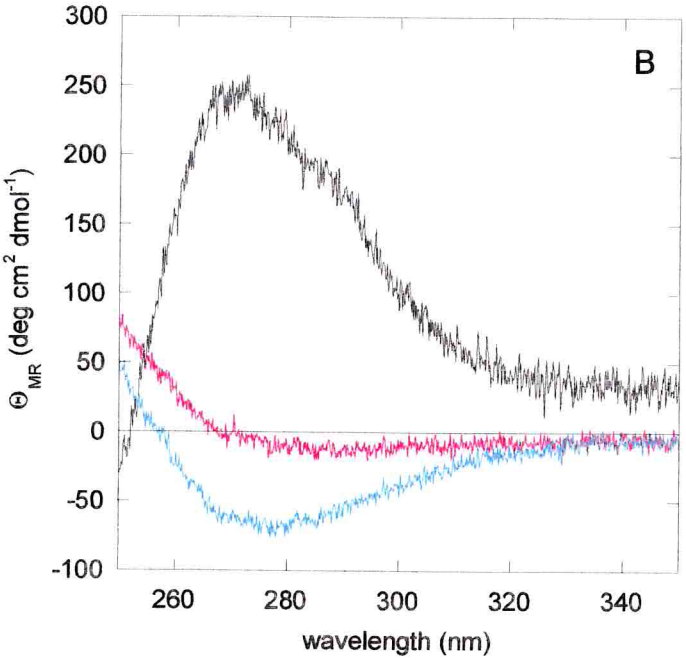


Figure 11: CD spectroscopy. CD spectra (mean residue ellipticity, θ_{MR}, deg cm^2 dmol^{-1}) in the far-UV (190-250 nm, A) of egf-14 (black), f5b (red), f6b (blue), f5a (orange) and f6a (light blue) and in the near-UV (250-350 nm, B) of egf-14 (black), f5b (red), f6b (blue).

4 Discussion

4.1 The "frame-shift" approach

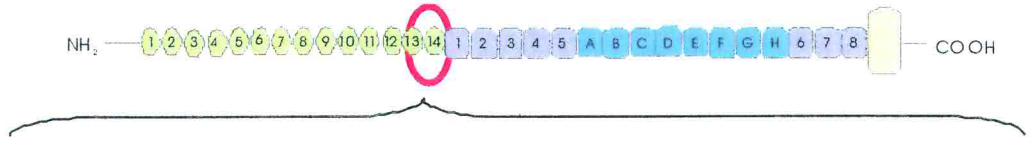
Proteins targeted to the extra-cellular environment can contain several tandem cysteine-rich domains, and the correct pairing of cysteines to form disulfide bridges is critical to reach the final native fold. In principle, two different factors can determine the pairing of cysteines to give disulfide bonds in multi-domain 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, 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. An incomplete EGF-like module 13 amino acids long and with 3 cysteine residues precedes the first EGF-like domain. Consequently, different frames are possible according to the cysteine chosen as the first (Fig.1). The tandem repeats have thus an "arbitrary" starting point that determines its folding frame. In principle the EGF-like module can be slid upward or downward obtaining different folding frames all composed of 33 amino acids modules containing 6 cysteines each. Nevertheless, there is a unique folding frame, which is determined by the spontaneous tendency of the modules to fold according to the EGF-like structure. 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 and sliding this window over the sequence of tenascin EGF repeats 13 and 14 (residues 560-622) by one cysteine at each step. The peptide sequences where choosen in order to minimize differences other than the cysteine spacing. We obtained six peptides that are all 33 residues long, contain six cysteines, and bear a partial overlap in the sequence (Fig. 2). The cases where a cysteine residue is located at the terminal position were avoided in order not to modify its reactivity. In peptide f3 it was necessary to add two extra amino acids. Alanine and

Serine were chosen to minimize physico-chemical perturbations. While the first peptide corresponds to the

Figure 1: Human tenascin contains 14 EGF-like repeats, 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 consequently, different frames are possible according to the cysteine chosen as the first. The correct folding frame is shown in red.

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 then intra-repeat disulfide bonds form. In other words, we forced misfolding to occur within short peptides that nevertheless maintain their native sequence.

Modules 13 and 14 were chose among the others mainly because EGF-like 14 was demonstrated to present biological activity (27). Moreover, the boundaries of the tandem repeats are defined at its C-terminal part by a fibronectin type-III module while, as mentioned before, at its N-terminus an incomplete EGF-like module introduces uncertainty in the experimental model.




```
EQRCPSDCHGQGRCVDGQCICHEGFTGLDCGQHSCPSDCNNLGQCVSGRCICNEGYSGEDCSE
                              GQHSCPSDCNNLGQCVSGRCICNEGYSGEDCSE   f1
                     EGFTGLDCGQHSCPSDCNNLGQCVSGRCICNEG            f2
                ICHEGFTGLDCGQHSCPSDCNNLGQCVSGRCI                  f3
           DGQCICHEGFTGLDCGQHSCPSDCNNLGQCVSG                      f4
      HGQGRCVDGQCICHEGFTGLDCGQHSCPSDCNN                           f5
 SDCHGQGRCVDGQCICHEGFTGLDCGQHSCPSD                                f6
```

Figure 2: Sequence of the peptide under study and their position into the EGF-13 and EGF-14 modules

4.2 Oxidative folding

Because the EGF repeat is one of the most commonly employed building block in extra-cellular proteins, we wondered if there might be a kinetic reason that largely favors the correct formation of disulfide bonds within the same EGF repeat, or in other words, if the EGF-type repeats are so successful in respect to all other possible arrangements of disulfide patterns because they are stable, fast folding modules. Experimental results at least partially support this hypothesis. Although the disappearance rates of the reduced frame-shifted peptides, including egf-14, are all within the same order of magnitude, the folding of egf-14 is indeed faster (Fig. 3). Moreover, 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 hours is the major species. Both kinetic and thermodynamic factors are therefore favoring the EGF-

like topology, determining a preferential "folding frame" in the cluster of highly repeated domains.

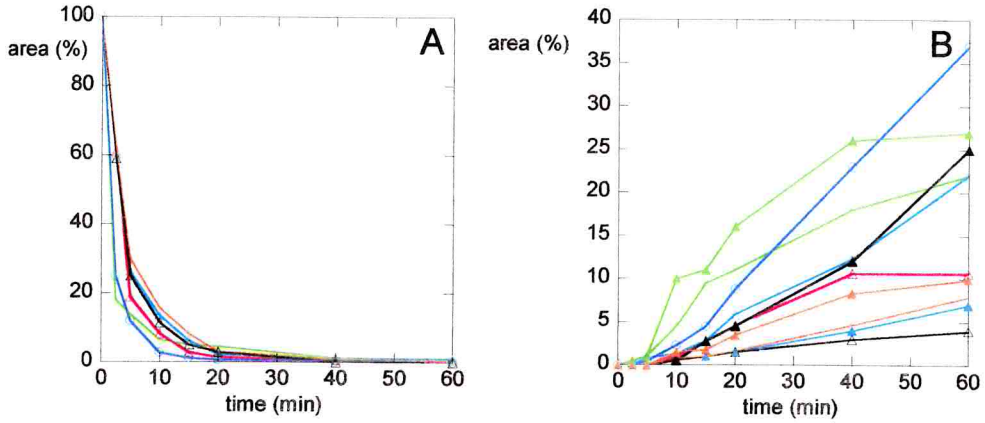


Figure 3: Oxidative folding kinetics. Panel A, disappearance of the starting product (%, area of the initial reduced form with respect to the total integrated area) for egf-14 (blue), f2 (green), f3 (red), f4 (light blue), f5 (black), f6 (orange). Panel B, formation of three-disulfide species (%, area of the 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) originating from the same peptide are shown as empty and filled triangles, respectively. Oxidative folding kinetics were followed by HPLC and UV detection at 214 nm

4.3 Peptide structure

Detailed structural studies of frame-shifted peptides have been hampered by the complexity of the mixtures obtained in the oxidative folding studies, and by the small quantities of three-disulfide species that could be recovered. Our efforts pointed towards the characterization of those products that represented a major species in the mixture, that were well separated in HPLC chromatograms, and that displayed a large difference between the retention time of the reduced and fully oxidized specie (f5a, f5b, f6a, f6b) (Fig 4). The latter was considered an important indication of effective burial of hydrophobic residues upon folding, with the formation of a relatively compact structure.

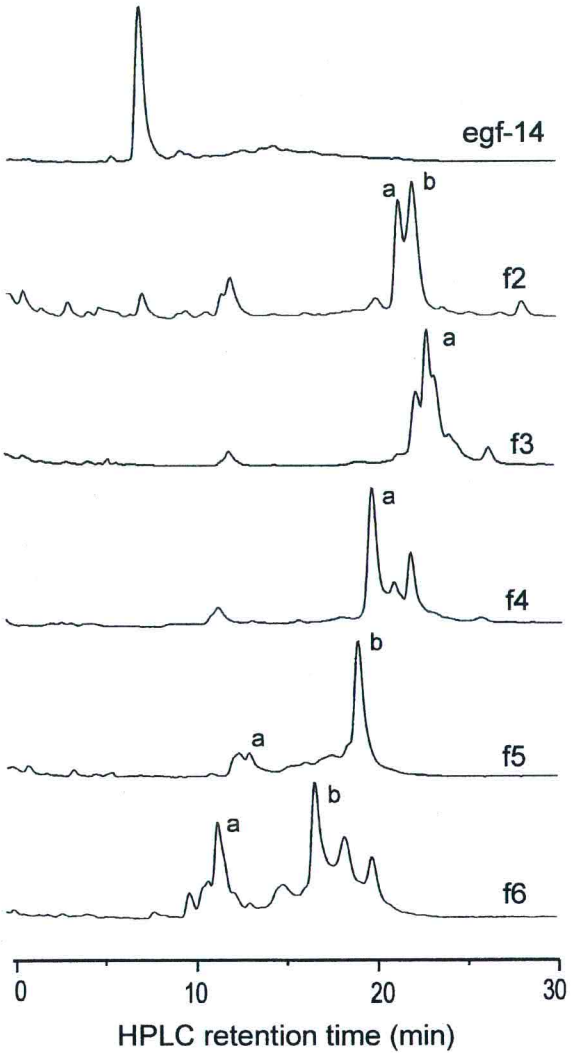


Figure 4: Equilibrium mixtures. HPLC profiles of the oxidative folding reactions after 24 h. Detection was performed at UV at 214 nm. The species isolated for structural studies are labeled with letters **a** and **b**.

Peptides with hydrophobic residues deeply buried into the structure expose to the external environment a much more hydrophilic surface. This improves the solubility in buffer A with a consequent reduction of the RP-HPLC retention time. A similar behavior 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 resulting in products with retention times similar to that of the totally reduced forms. 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. In

contrast, egf-14 displays the CD spectra characteristics of a compact globular domain. These results have been confirmed even by NMR studies (Fig. 5)(74).

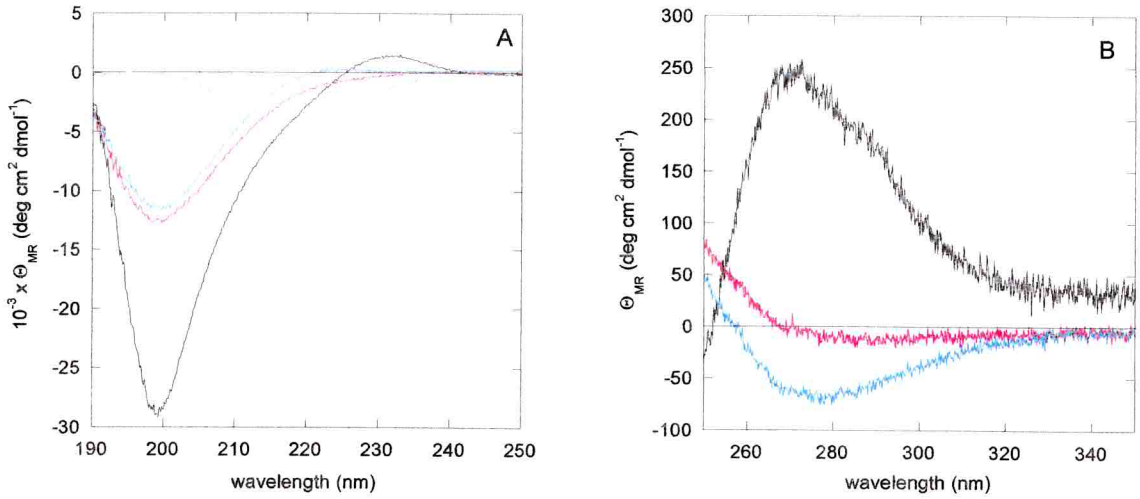


Figure 5: CD spectra (mean residue ellipticity, θ_{MR}, deg cm^2 dmol^{-1}) in the far-UV (190-250 nm, A) of egf-14 (black), f5b (red), f6b (blue), f5a (orange) and f6a (light blue) and in the near-UV (250-350 nm, B) of egf-14 (black), f5b (red), f6b (blue). CD studies suggest that the products of the oxidative folding of frame-shifted peptides (f5a, f5b, f6a, f6b) are highly flexible in solution and only partially structured. In contrast, egf-14 displays the CD spectra characteristics of a compact globular domain.

4.4 Relevance to folding in vivo

The folding *in vivo* of an extra-cellular protein containing disulfide rich domains, the low-density lipoprotein receptor, has shed new light on the folding process in the living cell (75). In contrast to the commonly assumed "vectorial" model, in which domains in a multi-domain 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 collapses 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, which operates in concerted action and in physical association with chaperone proteins (76-78). 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 PDI is still lacking, but homology modeling of the peptide recognition domain b' of PDI suggests that a small hydrophobic pocket capable of hosting even single amino acids could represent the binding site (79). In a similar way, a heptapeptide fragment of alternating hydrophobic residues has been shown to be recognized by BiP (80), 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 (81), 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 latters, 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.

4.5 Prospects and conclusions

The results obtained raise new questions and suggest further experiments. Important information could emerge from the study of peptides containing more EGF-like tandem repeats, the simplest of which is a double EGF-like module. Within this model the structure, stability and folding dynamics could be studied in a system that is closer to the reality. A similar investigation will probably require a different methodology for the peptide preparation. The chemical synthesis of a double EGF-like module peptide is not, in fact, a straightforward process because the length of the peptide (66 residues) and the presence of 12 cysteines can create difficulties during synthesis and purification. Molecular biology methodologies like the gene synthesis coupled with the *in vivo* expression are probably more suitable for this purpose.

Further investigations are necessary to determine the disulfide topology and structure of f5 and f6 three disulfide bond products. Our intent to achieve these data was hampered by the small amounts available. The comparison of the structure between the native egf-14 and f6a, b and f5a, b can explain the differences in the retention times observed by HPLC. The determination of the egf-14 structure by NMR spectroscopy is in progress.

5 Experimental techniques

5.1 Peptide Synthesis

5.1.1 Amide bond formation

The formation of an amide bond is the result of a nucleophilic attack of the α-amino group of one α-amino acid to the carboxyl group of another α-amino acid. This reaction has high activation energy and can take place only in very harsh conditions, which are incompatible with the presence of other functional groups. *In vivo*, the amino acids polymerization requires the involvement of catalysts, i.e. the ribosomal proteins. In ribosome-mediated synthesis, the addition of a new amino acid to the growing polypeptide is driven by the carboxyl group, which is involved in a highly activated esteric bond with the tRNA. The same strategy is used in vitro, through the activation of the carboxyl group to generate a reactive species, which allows the nucleophilic attack from the amino group and the formation of the amide bond at room temperature.

In order to obtain a linear peptide the α-amino terminus of the growing chain must react only with the C-terminus carboxyl group of the in-coming amino acid. It is therefore necessary to block all reactive groups not involved in the amide bound formation in order to prevent the formation of branched products. At the same time the amino group of any new amino acid and the carboxyl group of the growing polypeptide must be protected to avoid uncontrolled polymerization. Three different types of protecting groups are used. (i) The protection of the functional groups of the amino acids side chains prevents undesirable reactions that could lead to branched peptides. These groups are stable during all peptide synthesis steps and are finally removed to give the desired peptide. (ii) The protecting group for Nα-amino function. The Nα-amino group of the entering amino acid must be protected in order to avoid reactions with the activated carboxyl group of the same reagent. This would bring to a in solution polymerization of the activated reagent and introduction on the growing chain of an variable number of residues at the same

coupling reaction. This group must be removed for the next coupling step to take place. (iii) The last type of protecting group is that on the C-terminus of the peptide. It must be stable through all synthesis steps to be removed selectively if required.

5.1.2 Solid-phase peptide synthesis

In the Solid-Phase Peptide Synthesis (SPPS) approach, proposed by Merrifield in the 1963 (Nobel Price in chemistry 1984), the peptide is built up on an insoluble polymeric support (82). With respect to the synthesis in solution this strategy has numerous important advantages. The procedure is technically easier since the whole synthesis is carried out in a unique vessel. As the growing peptide is linked to an insoluble support, the isolation of synthetic intermediates is no more necessary because reagents and by-products are filtered away after each step by washing. In the solid-phase strategy the first amino acid C-terminus is anchored to the polymer, which is at the same time the C-terminal-protecting group. The most important outcome of the introduction of the solid-phase synthesis is a significant simplification that allowed the whole process to be automated.

SPPS has two main drawbacks largely overweighed, anyway, by the already mentioned advantages and today partly over-come by technical improvements. First: since the peptide is linked to the resin, a characterization of the synthetic intermediate is difficult and its purification impossible. The second is related with the deleted peptides formed because of non-quantitative amino acid couplings or Nα-deprotection. These contaminants are chemically and structurally strictly related to the target molecule and, in some cases, difficult to eliminate.

Many reactions are required to successfully obtain a synthetic peptide and the different steps of coupling and deprotection must proceed in the correct order until the whole synthesis is complete (Fig. 1). The polymerization proceeds from the C to the N-terminus which is blocked by a Nα-protection group. In order to add a new amino acid, after each coupling reaction, the next step is the removal of the Nα-blocking group. Once the Nα-terminus of the growing peptide is free, a new coupling reaction can be performed adding

the next activated side chain-protected amino acid. The polypeptide chain is now one amino acid longer. The newly introduced residue has its Nα-terminus still blocked and a Nα deprotection is required to proceed. The procedure is reiterated until all residues have been introduced, one at a time, in the correct order to finally achieve the desired peptide. The Nα deprotected product is then cleaved from the resin and side-chain deprotected in a unique reaction. The soluble deprotected peptide should now be purified from side products, mainly originated by not quantitative coupling reactions.

5.1.3 Solid supports

Chemical and physical properties of the solid carrier are essential in SPPS. First of all, the resin must be chemically inert to all reagents and solvents employed during the synthesis and cleavage-deprotection and at the same time must not interact with the peptide chain in any way. In addition, the mechanical stability of the polymer particles must permit an easy handling and a fast filtration. Most popular supports in use to date are polyethylene glycol-grafted polystyrene resins (83,84). These carriers have been demonstrated to give better performances than the cross-linked polystyrene resin originally used by Merrifield because their structure permits a better solvation of the peptide.

5.1.4 Side Chain protection and N-α protection

The two major protecting schemes used today in SPPS are the Boc and the Fmoc approaches. The denomination refers to the Nα protecting strategy. The t-Butoxycarbonyl (Boc) protecting group is removed by acidolysis and all the side-chain protecting groups are stable to the treatment with moderately strong acid solutions. The final treatment of

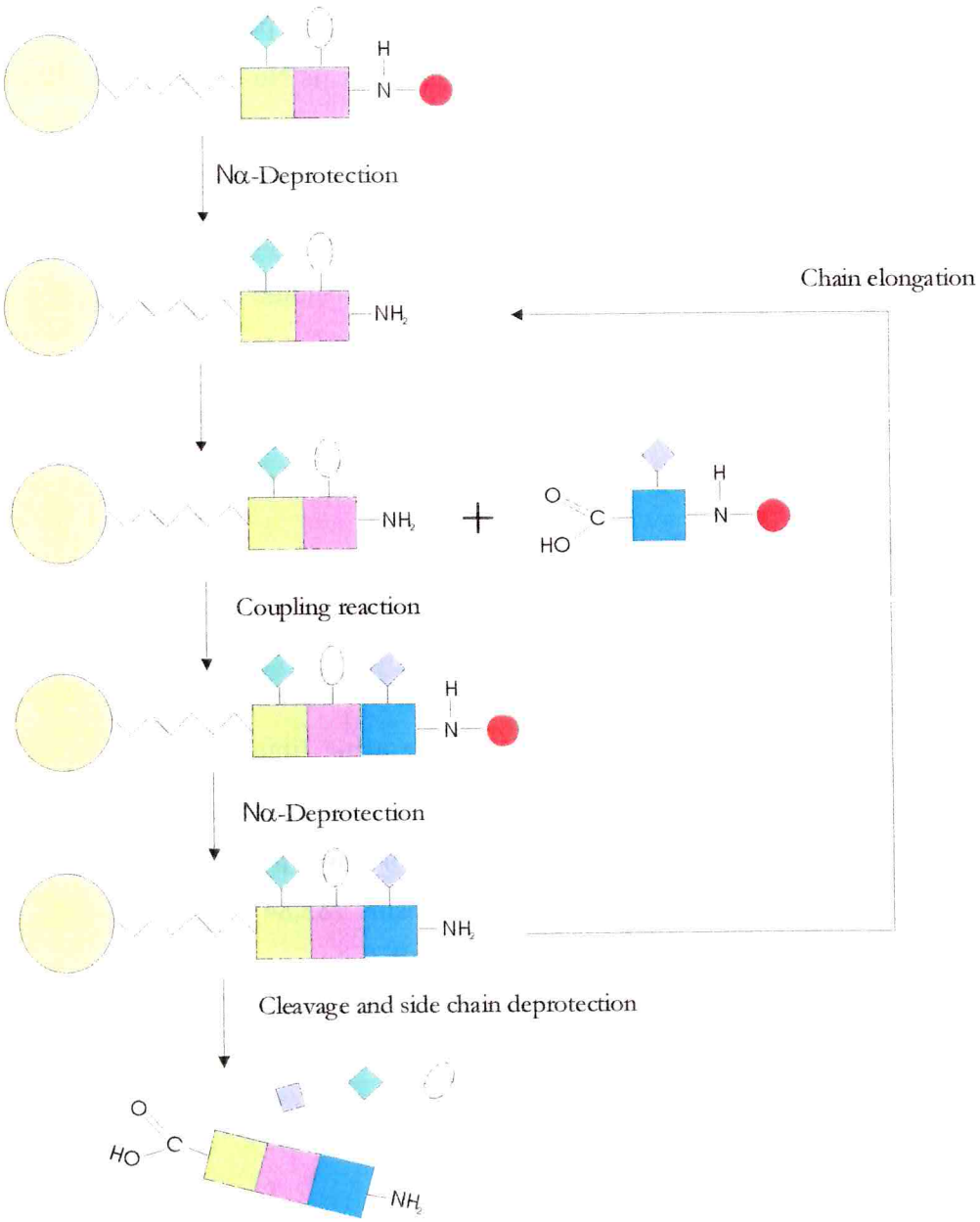


Figure 1: Schematic representation of the SPPS procedures. Amino acids are represented with squares while diamonds and ellipse are the side chain protecting groups. Red circle represents the Nα-protection. The growing peptide is shown linked to the insoluble support.

cleavage-deprotection must be carried out with strong acids, usually liquid hydrogen fluoride. The Fluorenylmethoxycarbonyl group (Fmoc) is labile in a solution of secondary amines but is stable in acidic conditions (85). The deprotection, carried out with a 20% (Vol/Vol) piperidine solution in DMF, takes few seconds to reach completion and is considered the standard procedure. The side chain protecting groups and peptide-resin linkage are labile to moderately strong acids. Usually, trifluoroacetic acid is used for this purpose. The Fmoc strategy was used in our work.

The role of Nα protecting groups is to suppress the nucleophilic reactivity of the amino group in these synthesis steps when it is not required. The protecting group, therefore, must be stable during the coupling reaction and should be easily removed before the next elongation step.

To avoid the branching of the peptide, all amino groups other than the Nα must be protected for the whole duration of the synthesis process. The ε-amino group of the lysine side chain is generally blocked with a tert-butoxycarbonyl (Boc) protecting group, which can be removed with TFA. Arginine could be used unprotected, because the strong basicity of its side-chain guanidino group makes it unreactive in normal conditions (86). Nevertheless, due to its poor solubility and side reactions, arginine is generally used side chain protected. 2,2,4,6,7-Pentamethyl-dihydrobenzofurane-5-sulphonyl (Pdf) is the most common protecting group.

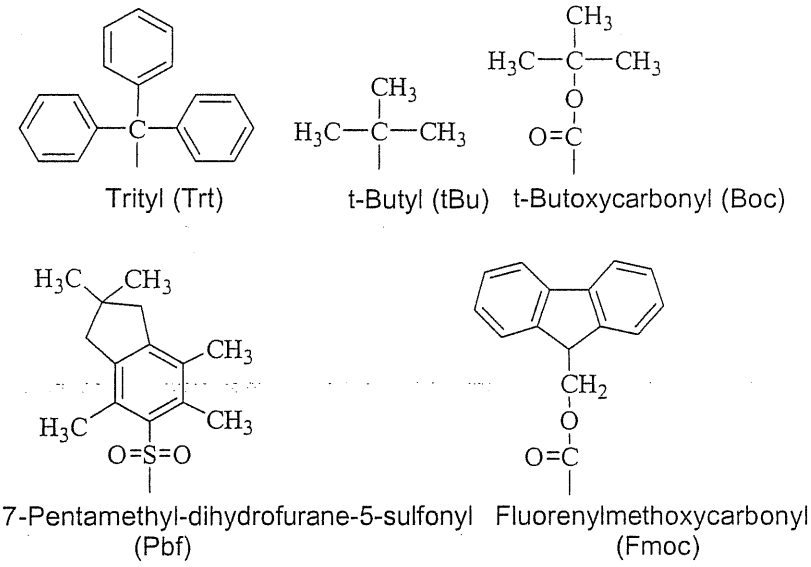


Figure 2: Structures of some of the most important side chain protecting groups. Fmoc is the standard Nα-protection in SPPS.

The protecting group ter-Butyl (tBut) for aspartic and glutamic acid side chains is largely employed in SPPS due to its stability during synthesis and its acid sensitivity, which makes it cleavable with TFA. The hydroxyl groups of Ser and Thr can react with acylating reagents and for this reason are used as side chain protected amino acids. Trifluoroacetic acid treatment is sufficient for deprotecting hydroxyl groups protected as tert-butyl ether. All other reacting groups on amino acid side chains must be blocked with suitable protecting groups. Cysteine thiols can react with amines during acylation reaction because they are good nucleophyles and can be even oxidized to give inter and intra-chain disulfide bonds. For this reason sulfhidryl groups must be kept blocked until the end of the synthesis or longer, if necessary, to avoid undesired disulfide formation. The most common used protecting group for Cys side chain is S-triphenylmethyl (Trt) which can be removed with trifloroacetic acid treatment.

5.1.5 Chain elongation:

Amino acid activation and coupling

The activation is necessary for peptide bond formation because carboxylic acids simply produce salts with amines at room temperature. Activation is achieved by attachment of an electron-withdrawing leaving group to the α-carboxylic function. Acylating species can be prepared separately and purified before being used in the coupling reaction or, in the case of activation reagents, formed *in situ* in the presence of the amino component.

Preformed active esters

Pentafluorophenyl esters are the most common representatives of the pre-formed reagents. They are less activated if compared with other acylating compounds and for this reason they generally cause less extensive side reactions, including racemization (87). Because of their under-activation, more time is needed for the reaction to reach completion.

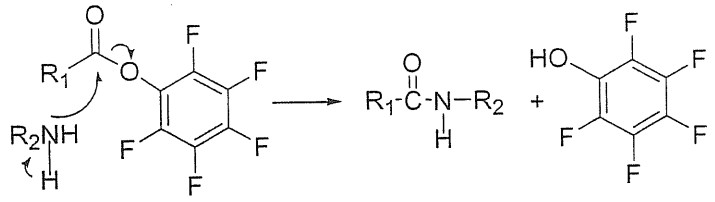


Figure 3: The pentafluorophenyl esters of the amino acids are low-activated acylating species. With these agents coupling reactions require longer time to go to completion.

Activation reagents

The most popular coupling strategy involves the use of activation reagents that create *in situ* the activated acylating species. The most significant features of these compounds are the simplicity and the rapidity of use. Couplings in fact are complete in one hour or less.

Carbodiimides

Dicyclohexylcarbodiimide (DCC) has been the most widely used activation reagent in peptide synthesis, both in solid and solution strategy, since its introduction in 1955 (88). Different pathways are thought to be involved in DCC activation mechanism but the most important involve the symmetrical anhydrides formation (Fig 4) (89). All of these pass through an *O*-acylisourea, which is an effective acylating agent (90,91). Several side reactions take place when DCC is used, such as dehydration of the side chain amide groups of Asn and Gln and racemization due to the potent activation (92). This problem can be relieved if an additive is used in the reaction. These chemicals, among which HOBt is the most common, interact with *O*-acylisourea or the symmetrical anhydrides giving a less reactive active ester. Acylation reaction mediated by this agent is still rapid and efficient but less prone to by-product formation.

The coupling reaction occurs in organic solvent at room temperature in presence of equimolecular amounts of carboxyl component and DCC.

Phosphonium and Uronium Reagents

Phophosphonium cations reacting with carboxylate anions generate acyloxyphosphonium species, which promote the peptide coupling rapidly and efficiently. 1H-benzotriazol-1-yloxy-tris(dimethylamino) phosphonium hexafluorophosphate (BOP), the first successful compound of this category, produces the hexamethylphosphorotiamide, side product that has been demonstrated to be highly toxic (93). This important drawback induced the development of safer related compounds. The most popular among the several phosphonium compounds available today, named PyBOP [1H-benzotriazol-1-yloxytris (pyrrolidino) phosphonium hexafluorophosphate], has the same activity of the parent molecule but forms less harmful side products.

Uronium salts are related molecules largely used in SPPS as coupling reagents with comparable performances in respect to phosphonium activators. The most common ones are *O*-(benzotriazol-1-yl)-1,1,3,3-tetramethyluronium tetraflouroborate (TBTU) and an aza-analog hexafluorophosphate salt known as HATU (94).

Phosphonium and uronium reagents follow similar reaction pathways in amino acid coupling (Fig.5). The coupling reagent is added in 1:1 ratio with the carboxyl species in an inert solvent, generally DMF or NMP, together with a tertiary amine to keep the carboxylic group in the anionic form.

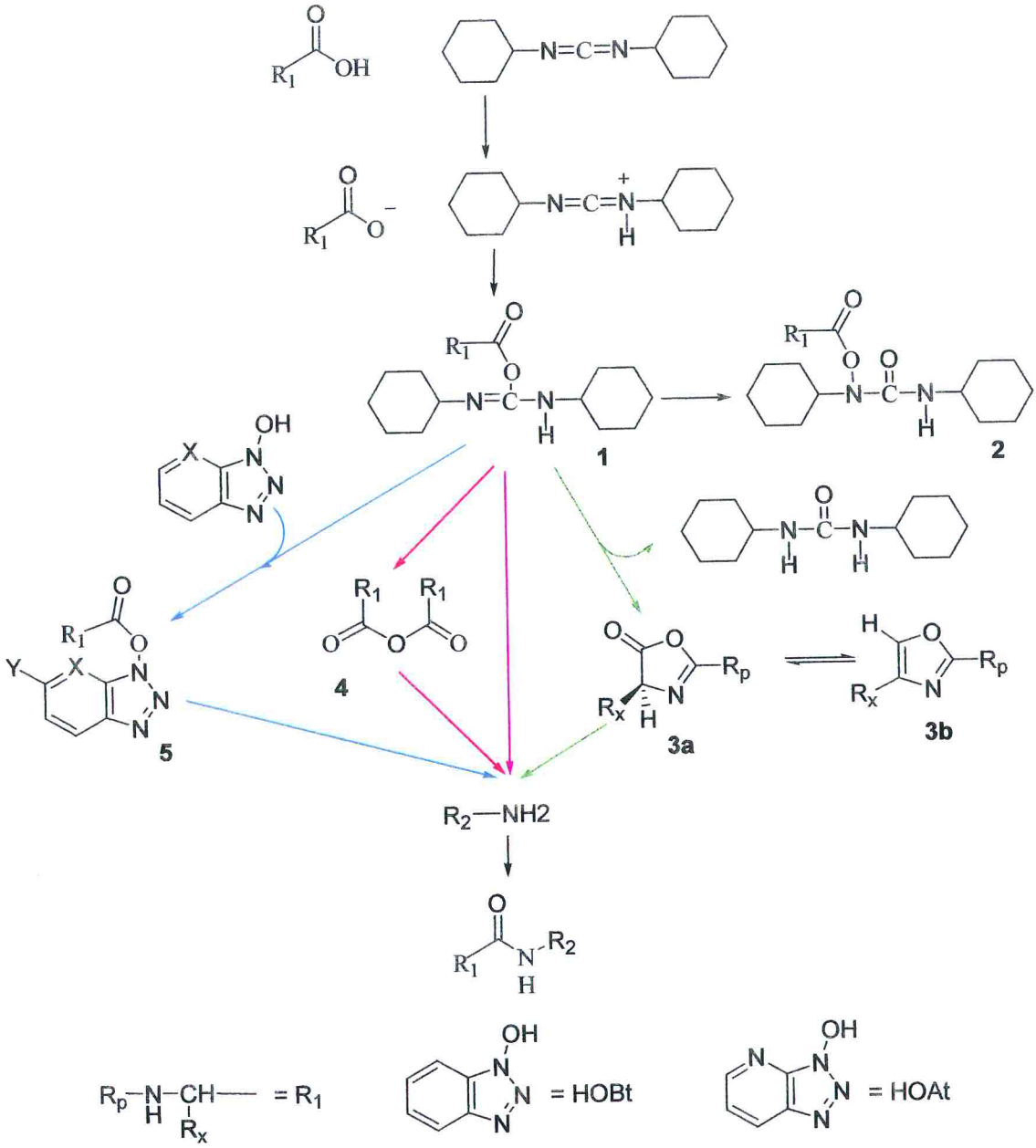


Figure 4: The mechanism of the carbodiimide activation starts with the formation of *O*-acylisourea (1). This is the most reactive specie that can attack the amino component to give the corresponding amide (Violet arrow). *O*-acylisourea can undergo a rearrangement to give the non-reactive specie N-acylurea (2). With a intra-molecular cyclization (Green arrow) *O*-acylisourea can give a 5(4H)-oxazolone (3), which can tautomarize with loss of chirality. When the reaction is carried out in solvents of low dielectric constant like CHCl3 or CH2Cl2 the specie 1 is instantaneously formed. If activation is performed in a more polar solvent like DMF, specie 2 and symmetrical anhydride (4) are formed (Red arrow). Additives like HOBt or HOAt are often used in DCC activation to reduce racemization. In presence of these substances the OBt active esters (5) are formed (Blue arrow). These species are less reactive than 1 but less prone to racemize.

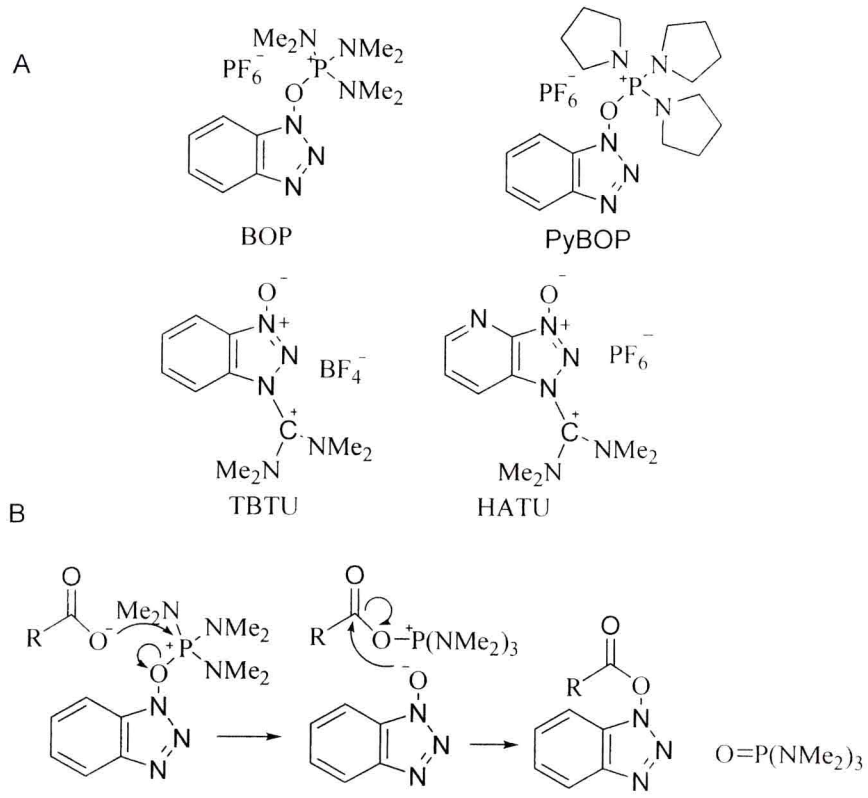


Figure 5: Panel A: The most important phosphonium and uronium coupling reagents are shown. **Panel B:** Mechanism of BOP-mediated coupling.

5.1.6 Cleavage and side chain deprotection

Acidolysis is commonly used to cleave peptides from the resin and, at the same time, remove the side chain protecting groups. Cleavage and deprotection of peptides obtained according to the Fmoc strategy can be achieved with trifloroacetic acid. Acid sensitive protecting groups are released in the medium as carbocations or other alkylating species and can react with sensitive unprotected amino acid side chains such as Trp, Met, Cys or Tyr. It is therefore necessary the presence of molecules, like anisol, thiols or silane derivatives in the cleavage mixture in order to trap the highly reactive carbocations. These chemicals, called scavengers, are present in large molar excess in respect to amino acid side chains and constitute the preferential target of alkylating agents, thus preserving the peptide from undesired modification.

5.1.7 Problems occurring during the synthesis

The most important problems are commonly related to the amino acid coupling reaction and the Nα deprotection step. The failure of these reactions is very often sequence dependent due to the fact that some side-chain protected sequences are prone to intermolecular aggregation or secondary structure formation. In these cases the N-terminal amino acid is less accessible to the reagents resulting in a significant decrease in the reaction yield. These difficulties can be addressed with different strategies that aim at disrupting the secondary structures. Most common are the use of a different solvent or a mix of solvents in order to change the chemical environment, or destabilize the aggregates increasing the reaction temperature. These strategies are not always successful and some peptides presenting particularly difficult sequences cannot be satisfactory synthesized by step-wise SPPS.

Different problems can arise from side reactions that may occur during the synthesis procedure depending on the peptide sequence and the employed methodology. The diketopiperazine formation is an intramolecular cyclization of the growing peptide that occurs at the stage of the third residue incorporation, considerably reducing the synthesis yield. The free amino group of the second residue can attack the peptide-resin linkage leading to the liberation of a cyclic dipeptide. This reaction is favored by the presence of good leaving groups in the peptide resin anchorage and by the nature of the first and second amino acid. Peptides where glycine or proline occur in these positions are particularly prone to diketopiperazine formation.

With the unique exception of glycine, in all other amino acids and proline the α-carbon is asymmetric, thus for each amino acid both R and S isomers are possible. Amino acids that occur in nature belong to the R group while cysteine, according to the roles of Chan, Ingold and Prelog, is an S amino acid. Non-conventional amino acids with the opposite configuration have been found in some proteins from plants, fungi and bacteria. The stereochemical fidelity of the coupling reaction is thus an essential requirement in peptide synthesis. Unfortunately, under standard condition of coupling some amino acids are prone to racemization. This side reaction heavily affects the incorporation of cysteine

residues where it can reach the unacceptable range of 5-33% expressed as the ratio of S:R of the obtained peptide (95).

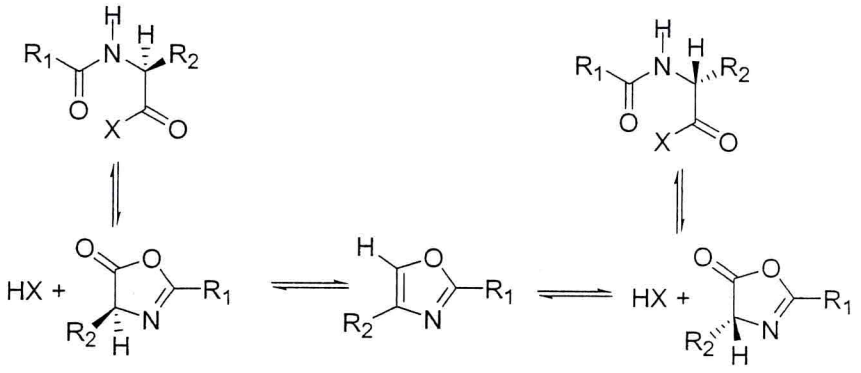


Figure 6: The azlactone formation/epimerization is promoted by acid activation and is influenced by the nature of group R_1. It is more rapid when R_1 is an alkyl, aryl or peptidyl group and is slower if R_1 a urethane or carbamate group like Boc or Fmoc Nα-protecting group.

There are two important racemization mechanisms: the base-induced abstraction of the proton bound to the α-carbon and the epimerization by azlactone formation (Fig 6). In the former case the re-protonation attack can take place on both sides of the planar ion resulting in the incorporation of a mixture of R and S amino acids in the peptide. The azlactone formation/epimerization is promoted by acid activation. Once an amino acid is correctly introduced further epimerization reactions are considered quite improbable. Considering that racemization is particularly enhanced in highly activated species and in presence of strong bases, the strategies to reduce it first involve the employment of a weaker base and less activated amino acids. Even the solvent plays its role in the process, and substituting pure DMF with the less polar CH_2Cl_2/DMF 1/1 mixture considerably reduces the racemization rate.

5.2 Mass spectrometry

5.2.1 Introduction

A mass spectrometer is an instrument that can separate ionized molecules according to their mass to charge ratio (m/z). The information provided by the output data is the molecular weight and the relative abundance of ions, but many other data about structure and purity of the parent molecule can be argued. The formation of ions in the gas phase is an essential step for the further processes of scanning and detection in a mass spectrometer. The sample is ionized in the ionization source where, generally, is introduced already in gas-phase. Recently, sources that can accept liquid samples have been developed; in this case ionization and vaporization take place contemporarily. Gas-phase ions are thus accelerated in the analyzer where they are separated according to the m/z ratio. A detector finally collects the signals, which are directly proportional to the corresponding abundance (96).

We can summarize the mass spectrometry analysis process in tree main steps:

1. Ionization
2. Scanning according to m/z ratio
3. Detection

5.2.2 Ionization

The traditional ionization technique in mass spectrometry is the electron impact (EI). In this source the target molecule in gaseous state is collided with accelerated electrons (70 MeV). The radical cation produced by the reaction breaks down to give a single charge ionized fragment. This methodology is highly informative when used for low-molecular weight organic molecules. Ions can in fact give precise information about the

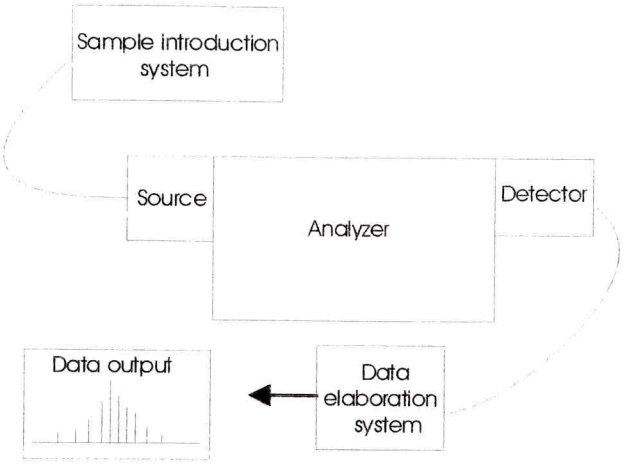


Figure 1: Schematic representation of the main parts of a mass-spectrometer

structure of the parent compound. Unfortunately, EI cannot be applied to not volatile molecules or on aqueous solutions, which are two common features of biological samples.

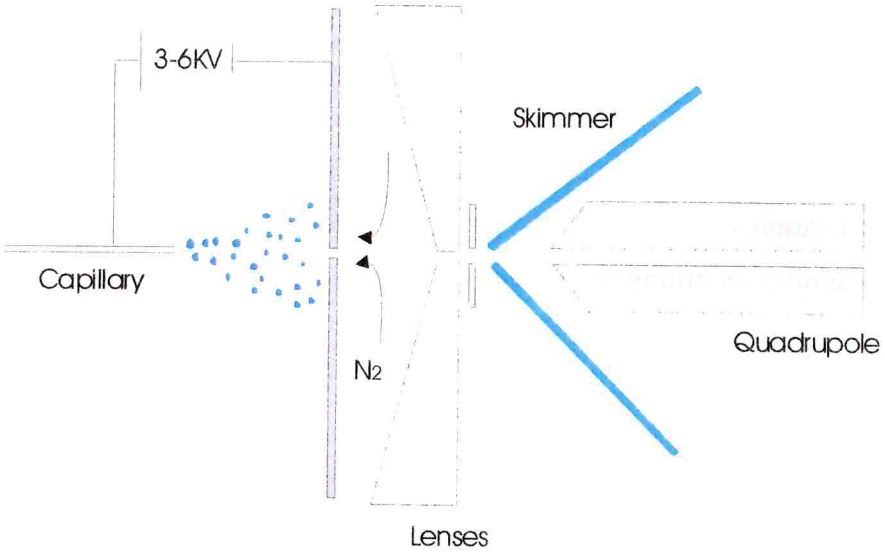


Figure 2: Sketch of the ion desolvation in electrospray process.

These limitations have been overcome by new, recently developed ionization sources, which gained a large popularity (97,98). Among all, electrospray ionization (ESI) is the most widely used. In the electrospray ionization (ESI) the analyte solution flows through a capillary in the source chamber. A strong electric field, approximately maintained at +5000 V, is generated between the capillary and the inlet of the scanning device. The

solvent ejected from the capillary is dispersed into a mist of highly charged droplets accelerated by the field towards the counter electrode. The droplet shrinking, due to the solvent evaporation, causes a strong electric repelling force that finally provokes the drop explosion. The cascade formation of smaller and smaller droplets repeats until the radius of curvature of the droplets becomes small enough for the field originated by the charge surface density to induce the desorption of the analyte molecules from the droplet to the gas phase. Desorped ionized analytes are suitable species for mass-spectrometry analysis. ESI is a soft ionization procedure that allows for the formation of charged ions without any fragmentation. Molecules with several ionization sites give multi-charged ions, but even molecules without ionizable groups can be detected, thanks to the formation of sodium, potassium or ammonium adducts. The production of multi-charged ions has important implications. Because mass spectrometry detects the m/z ratio, high molecular weight multi-charged molecules appear in the spectrum at m/z values that are fractions of the mass of the parent compound. This allows large molecules to give signals within the instrument range. A variety of algorithms have been developed to obtain the reconstruction of the original molecular weight starting from the spectra (99,100).

The spray formation in ESI can be enhanced by a gas jet (air or nitrogen) in order to accept larger solvent flows; in this case the source is named Ion Spray (IS).
The numerous advantages that ESI and IS offer in several applications, make them the methods of choice in different fields such as biochemistry, environmental chemistry, molecular medicine, drug and food analysis. ESI-MS can be efficiently applied to not volatile, ionic and polar compounds. Most of the molecules of biological interest such as DNA and polypeptides belong to this category of compounds. In many other applications like pharmacology and environmental science, ESI-MS can be used to directly analyze water-soluble molecules. With this technique it is no more necessary to go through derivatisation steps to increase the volatility of the compounds. The ESI is a mild ionization technique that limits the rate of fragmentation of the molecules. This feature lowers the number of information available, especially for small organic molecules, but on the other hand is very appreciated in the study of biological polymers. It is possible in this case to directly verify the molecular weight of large molecules. A it is no more

necessary to volatize the samples, electrospray ionization has emerged as a technique for studying in solution large and fragile molecules of biological interest. Spectra have been obtained of biopolymers having weights up to 130000 Dalton, allowing the use of the mass spectrometry analysis into the study of protein chemistry. The protein sample to be analyzed is typically a water or water-organic solvent solution with traces of acetic, formic or trifluoroacetic acid added to promote the sample ionization.

5.2.3 Quadrupole analyzer

The analyzer devices differ mainly in three characteristics;

1. Upper mass limit. It is the highest value of m/z ratio that can be measured by the device.
2. Transmission. It is the number of ions reaching the detector over that of ions produced in ionization process.
3. Resolution. Indicates the ability to distinguish two peaks with a small difference in m/z ratio.

The quadrupole is a low-resolution analyzer with a low mass limit (4000 Da) and it owes its popularity to the two main advantages it offers in respect to the other scanning devices, i.e. the low cost and the facility to combine it with different ionization sources. Quadrupoles are made up of four parallel rods with circular or hyperbolic section. Ions traveling through the analyzer are subjected to the combination of a constant electric field (DC) superimposed on a radiofrequency (RF). Equations of motion show that the ions trajectory, for given values of RF and DC, is dependent on the m/z ratio. Scanning is carried out with a systematic variation of DC and RF at uniform velocity. The values of RF and DC define whether ions of a particular mass-to-charge ratio, accelerated through the device, will follow a stable trajectory or not. At any moment only the selected m/z ratio is allowed to reach the analyzer; all other ions will drift apart and collide against the rods, not being detected (96). The resulting mass spectrum is a unique peak in case of single charged molecules or a series of peaks originated by the multiple charged compounds. From this series the mass is determined with a deconvolution performed by a

computer algorithm. Finally, a single peak is originated corresponding to the analyte mass.

5.2.4 Liquid chromatography-mass spectrometry

HPLC is a rapid procedure that permits to separate a large number of water-soluble compounds and high molecular weight molecules. The use of a mass spectrometer as a liquid chromatography detector offers higher sensitivity and much more specific information than any other kind of detector. Unfortunately, the coupling of these techniques was hampered for many years by several difficulties. On the one hand, HPLC works with liquid phases at high pressure; on the other hand mass spectrometer must be kept at a very low pressure for appropriate performance and can analyze only gaseous samples. This difficult integration was finally made possible by the introduction of new ionization sources; ESI and IS among all. The atmospheric pressure ionization (API) source is a device where the ionization is carried out at atmospheric pressure. Most common commercially available API sources rely on a IS ionization system. Ions produced in a vaporization chamber are continuously introduced in the spectrometer through a narrow inlet (orifice). The small diameter of this opening allows the vacuum system to maintain a correct pressure (10^{-5}-10^{-7} Torr) in the analyzer compartment. A nitrogen stream blown through the orifice protects the analyzer, sweeping away liquid chromatography vaporized solvents and not ionized molecules (101).

5.2.5 Mass spectrometry of proteins

In principle, all molecules that can be charged are accessible to ES-MS analysis. For many years peptides and proteins were excluded from mass spectrometry studies due to the difficulty to vaporize and ionize these molecules without destroying them. Since 1988, when two very different solutions to the problem were proposed, biochemists can

rely on the sensitivity and accuracy of mass spectrometry not only for molecular weight determination but even for innovative applications where the mass is already known (102,103). To date, mass spectrometry is the method of choice for the characterization of primary structure of proteins, to identify post-translational modifications and disulfide bond topology. Certain classes of proteins have proved to be very difficult to be analyzed, including certain insoluble membrane proteins, proteins with very stable tertiary structure and with a low number of basic residues. Water solubility of insoluble protein can be improved with detergents and salt buffers, which however have the drawback of interfering with the electrospray ionization and compete with the analyte molecules for charges.

Recently several developments have extended the use of MS to the study of weak non-covalent associations between biopolymers (104). The gentle electrospray ionization process, in appropriate conditions, allows a wide range of non-covalent complexes to pass intact to the gas phase and be analyzed by the mass spectrometer. The study of super-molecular interactions of proteins is nowadays the most innovative and challenging biological application of mass spectrometry.

5.3 Amino acid analysis

5.3.1 Introduction

An exact knowledge of protein quantities or composition is often a necessary premise for further studies. Amino acid analysis is a reliable methodology used to determine the amino acid composition and quantity of proteins and peptides. When the amino acid composition is already known, this procedure is a precise method for protein quantification. The relative amino acid composition gives also a characteristic profile of proteins, which is often sufficient for their identification. The methodology is based on a chromatographic separation of the free amino acid mixture obtained by total hydrolysis of the protein under study. The amino acids constituting the test sample are previously derivatized. Derivatization has the role to improve the absorbance of the free amino acids in order to make them easily detectable by ultraviolet-visible or fluorescence detectors.

Figure 3: Schematic representation of the amino acid analysis procedure.

In order to obtain accurate results, the amino acid analysis requires the use of highly purified samples and the employment of internal and external standards. For this purpose an accurately known amount of free amino acids solution is processed in parallel with the samples. The comparison between the internal standard and an identical not processed external standard permits an estimation of the general recovery of the amino acids.

5.3.2 Hydrolysis

Acid hydrolysis is the most used common methodology. The lyophilised sample is hydrolysed by HCl vapours under vacuo in the presence of a small amount of phenol. The reaction is carried out for 1 hour at 150°C or over night at 100 °C. Such harsh procedure causes a complete or partial destruction of several amino acids. Tryptophan is destroyed, serine and threonine are partially destroyed, methionine undergoes oxidation, cysteine is recovered as cystine and partially destroyed as well. Asparagine and glutamine are converted by deamination into aspartic and glutamic acid respectively. Those losses can be reduced by the application of an adeguate vacuum or by the introduction of an inert gas into the reaction chamber. Due to the loss of tryptophan, glutamine and asparagine, only 17 are the amino acids can be finally used for analysis.

5.3.3 Derivatization.

Most of the free amino acids cannot be detected by HPLC unless they have been derivatized. Derivatization is performed through the reaction of the free amino acids with phenylisothiocyanate under basic conditions in order to produce phenylthiocarbamyl amino acid derivatives. The reaction takes 20 minutes at room temperature and is followed by the lyophilization of the sample.

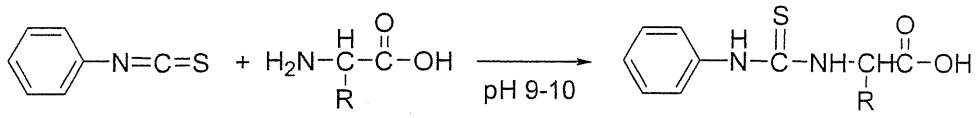


Figure 1: Reaction of the PITC derivative with the amino acid

5.3.4 HPLC separation.

The PTC-amino acids are separated by RP-HPLC using a triethyl ammonium acetate buffer system. The PTC chromophore is detected at 254 nm. Chromatographic peaks are identified using standard chromatograms as reference. Quantification of the amino acids is based on the peak areas that are proportional to the relative abundance of each amino acid.

5.4 Circular dichroism spectroscopy

5.4.1 Physical principles

Circular dichroism (CD) spectroscopy is a form of electron absorption spectroscopy that measures the difference in absorbance of right- and left-circularly polarized light by an optically active substance. This technique is very sensitive to the secondary structure of polypeptides and proteins particularly for spectra collected between 260 and 180 nm. The analysis of CD spectra can therefore provide valuable information about secondary structure of biological macromolecules.

Linear polarized light can be viewed as a superposition of opposite circular polarized light (clockwise and counter-clockwise) of equal amplitude and phase (Figure 1a). A projection of the two amplitudes according to the rules of vector addition, perpendicularly to the propagation direction, yields a line. When this light travels through an optically active sample with a different absorbance for the two components, the amplitude of the more absorbed component will be smaller than that of the less absorbed one. The consequence is that the superposition of the two components coming out from the sample is no longer a linearly polarized wave. The resulting field vector, in fact, does not oscillate along a line but it rotates along an ellipsoid path. Such a light wave is called an elliptically polarized light. The occurrence of ellipticity is due to differences in extinction coefficients ε of the sample for the two components (Fig.1b).
When a sample presents a difference in the refractive index n for the two circular components of light, their phases become different causing a rotation of the polarization plane (or the axes of the ellipse) by a small angle. This effect is called circular birefringence. It can be demonstrated that the two fenomena of ellipticity and birefringence are directly correlated and that when ellipticity exists, optical rotation must exist as well.

The difference in absorption of right and left components to be measured is always very small. The differential absorption is usually a few 1/100ths to a few 1/10th of a percent,

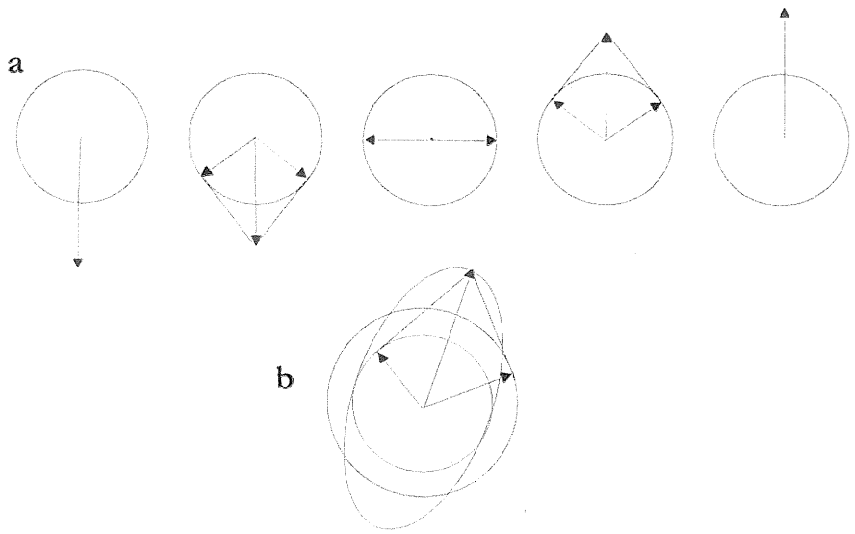


Figure 1: Linear polarized light can be viewed as a vector addition of opposite circular polarized light of equal amplitude and phase. A projection of the combined amplitudes perpendicular to the propagation direction yields a line (figure 1a). When this light passes through an optically active sample with a different extinction coefficient ε and a different refractive index n for the two components, the consequence is that a projection of the resulting amplitude now yields a tilted ellipse instead of a line (figure 1b).

but it can be determined quite accurately. Given a difference in absorption for the left and right circularly polarized light, in CD spectroscopy the Beer's law must be obeyed. The difference in absorption is given by:

$$\Delta A = \Delta \varepsilon l c$$

Where l is the light pathlength measured in centimeters and c the molar concentration. Despite the fact that CD is now measured as the difference in absorbance of right- and left- circularly polarized light as a function of wavelength, the unit ellipticity (θ) persists in CD measurements. Historically, ellipticity is the unit of circular dichroism and is defined as the tangent of the ratio of the minor to major elliptical axis. The two quantities are related by:

$$\theta = 32.98 \Delta A$$

To be able to compare ellipticity values of different samples we need to convert them into a normalized value. The most commonly used unit in protein and peptide work is the mean molar ellipticity $[\theta]$.

The relation between $[\theta]$ and the experimentally measured θ is:

$$[\theta]_{(\lambda)} = \frac{100\theta(\lambda)}{lc}$$

The units of $[\theta]$ are degrees cm^2 $dmol^{-1}$.

The relation between molar ellipticity and the difference in extinction coefficient is the following (105,106).

$$[\theta] = 3298\Delta\varepsilon$$

For CD measurements of proteins the mean residue molar ellipticity (MRE) is often used. It is defined as:

$$[\theta]_{MRE} = [\theta] / n$$

where n is the number of residues in the protein. MRE units are degrees cm^2 $dmol^{-1}$ $residue^{-1}$.

5.4.2 Circular dichroism spectroscopy of proteins and peptides

Circular dichroism is a sensitive technique for determining the structure of biopolymers. Intrinsically asymmetric chromophores or symmetric chromophores in asymmetric environments will interact differently with right- and left-circularly polarized light.

For proteins and peptides we mainly rely on the absorption of the peptide bonds (symmetric chromophores) and amino acid side chains in the ultraviolet region of the spectrum. The three aromatic side chains that occur in proteins (phenyl group of Phe, phenolic group of Tyr, and indole group of Trp) also have absorption bands in the ultraviolet spectrum. In proteins however, the contributions of these chromophores to the CD spectra in the far UV, where secondary structural information is located, is usually negligible. The disulfide group is an inherently asymmetric chromophore as it prefers a gauche conformation, and can lead to a broad CD absorption signal around 250 nm. CD

spectroscopy in the region of 230-190 nm is particularly powerful in monitoring conformational changes. In this region of the spectra it is possible to observe the effects of backbone conformational changes while CD effects at longer wavelengths (>230 nm) are due to aromatic side chains and prosthetic chromophore contributions.

All proteins with only α-helix as secondary structure produce spectra with a double minimum at 222 and 208-210 nm and a maximum at 191-193 nm. These features are characteristic of pure α-helix structures. All-β proteins have a single negative CD band between 210 and 225 nm and a single maximum between 190 and 200 nm. The signal intensities of β-sheet structures are much lower than those of α-helix. In proteins where both α-helix and β-sheet coexist the intensities of the signals reflect the amount of each secondary structure. Short polypeptides and proteins with secondary structures not defined as α or β usually show a strong negative signal around 200 nm.

5.4 3 Sample preparation and measurement

The protein solution used for measurement should contain only chemicals necessary for protein stability or solubility at the lowest possible concentration. Any compound which absorbs in the region of detection (250-190 nm) should be avoided and even buffers must be as transparent as possible in the far-UV. A typical buffer used in CD experiments is 10 mM phosphate, although low concentrations of Tris, perchlorate or borate are also acceptable. Potassium fluoride is preferred to NaCl to increase the ionic strength as the chloride ion has a strong UV absorbance at low wavelengths. A high pure protein sample is also required because any contaminating peptide will contribute to the CD signal. The

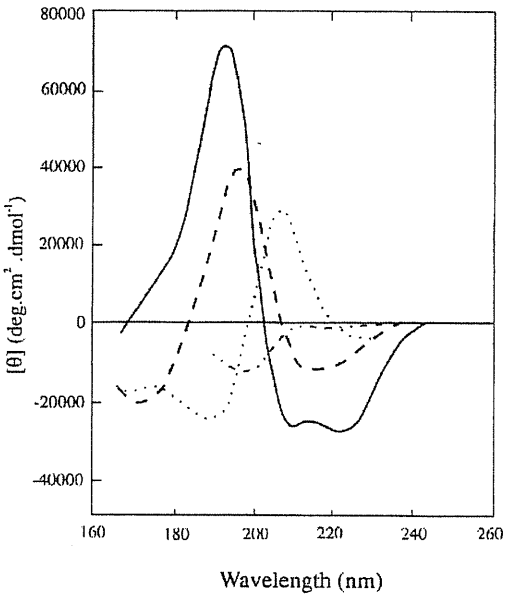


Figure 2: Circular dichroism spectra of "pure" secondary structures. Redrawn from Brahms & Brahms, 1980. Pure α helix spectrum is shown in solid line. Dashed line represents the pure β sheet and dotted line the spectrum due to the turn structures. The dash-dotted line shows the contribution of different structures.

concentration of peptide is an important aspect of the experiment. As a general rule, the total absorbance of the cell, buffer, and protein should be between 0.4 and 1.0. If the absorbance of the sample rises above 1.0, accurate CD measurements are not possible. Another consideration is that oxygen absorbs strongly below about 200 nm so very extensive purging with pure oxygen-free nitrogen is necessary for these measurements.

In a CD spectropolarimeter the light beam is originated by a xenon lamp and then filtered by a monochromator. A modulator induces a periodic variation of the light polarization. The light beam passes through all ellipticities from left circular to right circular passing through ellipticals and linear light. After interaction with the sample, the transmitted light is detected by a photomultiplier. Successive detections are performed at various wavelengths to generate a complete CD spectrum.

5.4.4 Methods to analyze protein conformation

Estimation of the secondary structure of a protein from its CD spectrum remains an empirical task despite the numerous proposed methods of analysis. All these methodologies are based on the same fundamental assumptions:

a. Contributions of individual secondary structural elements on the overall CD spectrum are additive and the effect of tertiary structure is negligible.
b. Only peptide chromophores are responsible for the CD spectrum.
c. Each structural element such as α-helix and β-sheet can be described by a single CD spectrum.

The simplest method of extracting secondary structure content from CD data is to assume that a spectrum is a linear combination of CD spectra of each contributing secondary structure type ("pure" α-helix, "pure" β-strand etc.) weighted by its abundance in the polypeptide conformation. The calculation is based on standard synthetic peptides used as models of pure secondary structures to obtain reference spectra. The CD spectra associated with pure secondary structures are shown in figure 2 (105,106).

Many different methods to extract protein conformation from CD spectra have been developed and most of them are available as web facilities provided by different research centers. As previously said all these methods are based on the assumption that the experimental spectra can be considered as linear combinations of the spectra of the pure secondary structural motifs. The accuracy of the different methodologies was compared by performing the extraction of the secondary structure of the same set of proteins . All the 16 proteins constituting the standard have a known structure solved by X-ray diffraction. All methods compared gave reliable results in α-helix prediction but their results are less consistent in the determination of the β-sheet and β-turn contents.

The most important methodologies are briefly described below.

Multilinear regression.

It is the oldest and simplest method of analyzes. It fits the experimental data with the standard spectra with the method of the least squares. MLR program is based on non-

constrained least-squares analysis. This method is the only that can be applied when the sample concentration is not precisely known. It provides good results in α-helix and β-sheet estimation but very poor results in the calculation of β-turn content. On the other hand the constrained least-squares analysis is used by LINCOMB that improves the estimation of the β-turn content (107).

Nowadays more recent programs that can provide, in most of the cases, more reliable structure estimations have overcome the least squares based programs.

Selection method.

In certain cases unreliable results are due to the presence in the reference standard of proteins with unusual CD spectra. Aromatic amino acids, disulfide bridges or uncommon conformations may originate these differences. The selection method aims to refer the experimental spectrum with a more suitable standard taking in consideration the proteins with the spectral characteristics most similar to the protein under study. Several available programs are based on different selection procedures. The CONTIN software, for instance, refers to the ridge regression analysis. Others selection methodologies are the variable selection (CDSSTR) (108) and the self-consistent method (SELCON software) (109) while K2D (110) software is based on a neural network program.

Appendix A

EGF-like f2

Characteristics

Sequence:
H-HEGFTGLDCGQHSCPSDCNNLGQCVSGRCICNE-OH
Amino acids: 33
Molecular weight 3451.2

Synthesis

Resin: TentaGel-S-PHB-Glu(t-Bu)Fmoc
Resin substitution: 0.2 mmol/g
Resin quantity: 0.5 g
Synthesis scale; 0.01 mmol
Solvent: DMF
Deprotection: 20% (V/V) Piperidine, 0.1M HOBt in DMF. 2 X 5 min.

AA position	Building Block	Activator	Number of couplings	Molar excess	Minutes
2	Fmoc-Asn(Trt)-OH	TBTU	1	4	60
3	Fmoc-Cys(Trt)-OPfp		1	4	120
	Fmoc-Cys(Trt)-OH	TBTU	1	4	120
Capping Reaction					
4	Fmoc-Ile-OH	TBTU	1	4	60
5	Fmoc-Cys(Trt)-OPfp		1	4	120
	Fmoc-Cys(Trt)-OH	TBTU	1	4	120
6	Fmoc-Arg(Pbf)-OH	TBTU	1	4	60
7	Fmoc-Gly-OH	TBTU	1	4	90
8	Fmoc-Ser(tBu)-OH	TBTU	1	4	60
9	Fmoc-Val-OH	TBTU	1	4	60
10	Fmoc-Cys(Trt)-OPfp		1	4	120
	Fmoc-Cys(Trt)-OH	TBTU	1	4	120
11	Fmoc-Gln(Trt)-OH	TBTU	1	4	90
12	Fmoc-Gly-OH	TBTU	1	4	90
13	Fmoc-Leu-OH	TBTU	1	4	60

	Fmoc-Leu-OH	PyBop	1	4	30
14	Fmoc-Asn(Trt)-OH	TBTU	1	4	90
15	Fmoc-Asn(Trt)-OH	TBTU	1	4	120
16	Fmoc-Cys(Trt)-OPfp		1	4	120
	Fmoc-Cys(Trt)-OH	TBTU	1	4	120
17	Fmoc-Asp(OtBu)-OH	TBTU	1	4	90
18	Fmoc-Ser(tBu)-OH	TBTU	1	4	60
19	Fmoc-Pro-OH	TBTU	1	4	60
20	Fmoc-Cys(Trt)-OPfp		1	4	120
	Fmoc-Cys(Trt)-OH	TBTU	1	4	120
21	Fmoc-Ser(tBu)-OH	TBTU	1	4	60
22	Fmoc-His(Trt)-OH	TBTU	1	4	60
	Fmoc-His(Trt)-OH	PyBop	1	4	60
23	Fmoc-Gln(Trt)-OH	TBTU	1	4	90
24	Fmoc-Gly-OH	TBTU	1	4	90
25	Fmoc-Cys(Trt)-OPfp		1	4	120
	Fmoc-Cys(Trt)-OH	TBTU	1	4	120
26	Fmoc-Asp(OtBu)-OH	TBTU	1	4	90
	Fmoc-Asp(OtBu)-OH	PyBop	1	4	60
27	Fmoc-Leu-OH	TBTU	1	4	90
28	Fmoc-Gly-OH	TBTU	1	4	90
29	Fmoc-Thr(tBu)-OH	TBTU	1	4	120
30	Fmoc-Phe-OH	TBTU	1	4	120
31	Fmoc-Gly-OH	TBTU	1	4	90
32	Fmoc-Glu(OtBu)-OH	TBTU	1	4	90
33	Fmoc-His(Trt)-OH	PyBop	1	4	120

EGF-like f3

Characteristics

Sequence:
H-ACHEGFTGLDCGQHSCPSDCNNLGQCVSGRCIS-OH
Amino acids: 33
Molecular weight 3398.2

Synthesis

Resin: TentaGel-S-PHB-Ser(t-Bu)Fmoc
Resin substitution: 0.2 mmol/g
Resin quantity: 0.5 g
Synthesis scale; 0.01 mmol
Solvent: DMF
Deprotection: 20% (V/V) Piperidine, 0.1M HOBt in DMF. 2 X 5 min.

AA position	Building Block	Activator	Number of couplings	Molar excess	Minutes
2	Fmoc-Ile-OH	TBTU	1	4	90
3	Fmoc-Cys(Trt)-OPfp		1	4	120
	Fmoc-Cys(Trt)-OH	TBTU	1	4	120
Capping Reaction					
4	Fmoc-Arg(Pbf)-OH	TBTU	1	4	120
5	Fmoc-Gly-OH	TBTU	1	4	60
6	Fmoc-Ser(tBu)-OH	TBTU	1	4	120
7	Fmoc-Val-OH	TBTU	1	4	90
8	Fmoc-Cys(Trt)-OPfp		1	4	120
	Fmoc-Cys(Trt)-OH	TBTU	1	4	120
9	Fmoc-Gln(Trt)-OH	TBTU	1	4	120
	Fmoc-Gln(Trt)-OH	PyBop	1	4	120
10	Fmoc-Gly-OH	TBTU	1	4	90
11	Fmoc-Leu-OH	TBTU	1	4	90
12	Fmoc-Asn(Trt)-OH	TBTU	1	4	90
13	Fmoc-Asn(Trt)-OH	TBTU	1	4	120
14	Fmoc-Cys(Trt)-OPfp		1	4	120
	Fmoc-Cys(Trt)-OH	TBTU	1	4	120
15	Fmoc-Asp(OtBu)-OH	TBTU	1	4	90
16	Fmoc-Ser(tBu)-OH	TBTU	1	4	60
17	Fmoc-Pro-OH	TBTU	1	4	60

20	Fmoc-Cys(Trt)-OPfp		1	4	120
	Fmoc-Cys(Trt)-OH	TBTU	1	4	120
21	Fmoc-Ser(tBu)-OH	TBTU	1	4	60
22	Fmoc-His(Trt)-OH	TBTU	1	4	90
	Fmoc-His(Trt)-OH	PyBop	1	4	60
23	Fmoc-Gln(Trt)-OH	TBTU	1	4	90
24	Fmoc-Gly-OH	TBTU	1	4	90
25	Fmoc-Cys(Trt)-OPfp		1	4	120
	Fmoc-Cys(Trt)-OH	TBTU	1	4	120
26	Fmoc-Asp(OtBu)-OH	TBTU	1	4	90
	Fmoc-Asp(OtBu)-OH	PyBop	1	4	60
27	Fmoc-Leu-OH	TBTU	1	4	120
	Fmoc-Leu-OH	PyBop	1	4	90
28	Fmoc-Gly-OH	TBTU	1	4	90
	Fmoc-Gly-OH	PyBop	1	4	90
29	Fmoc-Thr(tBu)-OH	TBTU	1	4	120
	Fmoc-Thr(tBu)-OH	PyBop	1	4	120
30	Fmoc-Phe-OH	TBTU	1	4	90
31	Fmoc-Gly-OH	TBTU	1	4	90
32	Fmoc-Glu(OtBu)-OH	TBTU	1	4	90
33	Fmoc-His(Trt)-OH	TBTU	1	4	120
	Fmoc-His(Trt)-OH	HATU	1	4	90
34	Fmoc-Cys(Trt)-OPfp		1	4	120
	Fmoc-Cys(Trt)-OH	TBTU	1	4	120
35	Fmoc-Ile-OH	TBTU	1	4	90
33	Fmoc-Ser(tBu)-OH	TBTU	1	4	90

EGF-like f4

Characteristics

Sequence:
H-DGQCICHEGFTGLDCGQHSCPSDCNNLGQCVS-OH
Amino acids: 32
Molecular weight 3327.3

Synthesis

Resin: TentaGel-S-PHB-Ser(t-Bu)Fmoc
Resin substitution: 0.2 mmol/g
Resin quantity: 0.5 g
Synthesis scale; 0.01 mmol
Solvent: DMF
Deprotection: 20% (V/V) Piperidine, 0.1M HOBt in DMF. 2 X 5 min.

AA position	Building Block	Activator	Number of couplings	Molar excess	Minutes
2	Fmoc-Val-OH	TBTU	1	4	90
3	Fmoc-Cys(Trt)-OPfp		1	4	120
	Fmoc-Cys(Trt)-OH	TBTU	1	4	120
CAPPING REACTION					
4	Fmoc-Gln(Trt)-OH	TBTU	1	4	120
5	Fmoc-Gly-OH	TBTU	1	4	120
6	Fmoc-Leu-OH	TBTU	1	4	120
7	Fmoc-Asn(Trt)-OH	TBTU	1	4	90
8	Fmoc-Asn(Trt)-OH	TBTU	1	4	120
9	Fmoc-Cys(Trt)-OPfp		1	4	120
	Fmoc-Cys(Trt)-OH	TBTU	1	4	120
10	Fmoc-Asp(OtBu)-OH	TBTU	1	4	90
11	Fmoc-Ser(tBu)-OH	TBTU	1	4	90
12	Fmoc-Pro-OH	TBTU	1	4	60
13	Fmoc-Cys(Trt)-OPfp		1	4	120
	Fmoc-Cys(Trt)-OH	TBTU	1	4	120
14	Fmoc-Ser(tBu)-OH	TBTU	1	4	90
15	Fmoc-His(Trt)-OH	TBTU	1	4	90
16	Fmoc-Gln(Trt)-OH	TBTU	1	4	90
17	Fmoc-Gly-OH	TBTU	1	4	90
18	Fmoc-Cys(Trt)-OPfp		1	4	120

	Fmoc-Cys(Trt)-OH	TBTU	1	4	120
19	Fmoc-Asp(OtBu)-OH	TBTU	1	4	90
	Fmoc-Asp(OtBu)-OH	PyBop	1	4	60
20	Fmoc-Leu-OH	TBTU	1	4	120
21	Fmoc-Gly-OH	TBTU	1	4	90
22	Fmoc-Thr(tBu)-OH	TBTU	1	4	120
	Fmoc-Thr(tBu)-OH	PyBop	1	4	90
23	Fmoc-Phe-OH	TBTU	1	4	90
24	Fmoc-Gly-OH	TBTU	1	4	90
	Fmoc-Gly-OH	PyBop	1	4	60
25	Fmoc-Glu(OtBu)-OH	TBTU	1	4	90
26	Fmoc-His(Trt)-OH	TBTU	1	4	120
	Fmoc-His(Trt)-OH	PyBop	1	4	90
27	Fmoc-Cys(Trt)-OPfp		1	4	120
	Fmoc-Cys(Trt)-OH	HATU	1	4	120
	Fmoc-Cys(Trt)-OH	TBTU	2	4	60-o/n
28	Fmoc-Ile-OH	TBTU	1	4	90
29	Fmoc-Cys(Trt)-OPfp		1	4	120
	Fmoc-Cys(Trt)-OH	PyBop	1	4	120
30	Fmoc-Gln(Trt)-OH	TBTU	1	4	90
31	Fmoc-Gly-OH	TBTU	1	4	90
32	Fmoc-Asp(OtBu)-OH	TBTU	1	4	120

EGF-like f5

Characteristics

Sequence:
H-DGQCICHEGFTGLDCGQHSCPSDCNNLGQCVS-OH
Amino acids: 33
Molecular weight 3477.3

Synthesis

Resin: TentaGel-S-PHB-Asn(t-Bu)Fmoc
Resin substitution: 0.2 mmol/g
Resin quantity: 0.5 g
Synthesis scale; 0.01 mmol
Solvent: DMF
Deprotection: 20% (V/V) Piperidine, 0.1M HOBt in DMF. 2 X 5 min.

AA position	Building Block	Activator	Number of couplings	Molar excess	Minutes
2	Fmoc-Asn(Trt)-OH	TBTU	1	4	120
3	Fmoc-Cys(Trt)-OPfp		1	4	120
	Fmoc-Cys(Trt)-OH	TBTU	1	4	120
CAPPING REACTION					
4	Fmoc-Asp(OtBu)-OH	TBTU	1	4	90
5	Fmoc-Ser(tBu)-OH	TBTU	1	4	90
6	Fmoc-Pro-OH	TBTU	1	4	90
7	Fmoc-Cys(Trt)-OPfp		1	4	120
	Fmoc-Cys(Trt)-OH	TBTU	1	4	120
8	Fmoc-Ser(tBu)-OH	TBTU	1	4	90
9	Fmoc-His(Trt)-OH	TBTU	1	4	120
	Fmoc-His(Trt)-OH	PyBop	1	4	120
10	Fmoc-Gln(Trt)-OH	TBTU	1	4	90
11	Fmoc-Gly-OH	TBTU	1	4	90
12	Fmoc-Cys(Trt)-OPfp		1	4	120
	Fmoc-Cys(Trt)-OH	TBTU	1	4	120
13	Fmoc-Asp(OtBu)-OH	TBTU	1	4	90
14	Fmoc-Leu-OH	TBTU	1	4	120
15	Fmoc-Gly-OH	TBTU	1	4	90
16	Fmoc-Thr(tBu)-OH	TBTU	1	4	120

	Fmoc-Thr(tBu)-OH	PyBop	1	4	90
17	Fmoc-Phe-OH	TBTU	1	4	90
18	Fmoc-Gly-OH	TBTU	1	4	90
19	Fmoc-Glu(OtBu)-OH	TBTU	1	4	90
20	Fmoc-His(Trt)-OH	TBTU	1	4	120
	Fmoc-His(Trt)-OH	PyBop	1	4	90
21	Fmoc-Cys(Trt)-OPfp		1	4	120
	Fmoc-Cys(Trt)-OH	HATU	1	4	120
22	Fmoc-Ile-OH	TBTU	1	4	90
	Fmoc-Ile-OH	PyBop	1	4	90
23	Fmoc-Cys(Trt)-OPfp		1	4	120
	Fmoc-Cys(Trt)-OH	TBTU	1	4	120
24	Fmoc-Gln(Trt)-OH	TBTU	1	4	90
	Fmoc-Gln(Trt)-OH	PyBop	1	4	90
25	Fmoc-Gly-OH	TBTU	1	4	90
	Fmoc-Gly-OH	PyBop	1	4	90
26	Fmoc-Asp(OtBu)-OH	TBTU	1	4	90
27	Fmoc-Val-OH	TBTU	1	4	120
28	Fmoc-Cys(Trt)-OPfp		1	4	120
	Fmoc-Cys(Trt)-OH	HATU	3	4	120
	Fmoc-Cys(Trt)-OH	PyBop	2	4	Over/night
29	Fmoc-Arg(Pbf)-OH	TBTU	1	4	90
	Fmoc-Arg(Pbf)-OH	PyBop	1	4	Over/night
30	Fmoc-Gly-OH	TBTU	1	4	90
31	Fmoc-Gln(Trt)-OH	TBTU	1	4	90
	Fmoc-Gln(Trt)-OH	PyBop	1	4	90
32	Fmoc-Gly-OH	TBTU	1	4	90
33	Fmoc-His(Trt)-OH	TBTU	1	4	120

EGF-like f6

Characteristics

Sequence:
H-SDCHGQGRCVDGQCICHEGFTGLDCGQHSCPSD-OH
Amino acids: 33
Molecular weight 3451.3

Synthesis

Resin: TentaGel-S-PHB-AsP(t-Bu)Fmoc
Resin substitution: 0.18 mmol/g
Resin quantity: 0.55g
Synthesis scale; 0.01 mmo
Solvent: DMF
Deprotection: 20% (V/V) Piperidine, 0.1M HOBt in DMF. 2 X 5 min.

AA position	Building Block	Activator	Number of couplings	Molar excess	Minutes
2	Fmoc-Asp(OtBu)-OH	TBTU	1	4	90
3	Fmoc-Ser(tBu)-OH	TBTU	1	4	90
4	Fmoc-Pro-OH	TBTU	1	4	90
5	Fmoc-Cys(Trt)-OPfp		1	4	120
	Fmoc-Cys(Trt)-OH	TBTU	1	4	120
CAPPING REACTION					
5	Fmoc-Ser(tBu)-OH	TBTU	1	4	90
6	Fmoc-His(Trt)-OH	TBTU	1	4	90
7	Fmoc-Gln(Trt)-OH	TBTU	1	4	60
8	Fmoc-Gly-OH	TBTU	1	4	90
9	Fmoc-Cys(Trt)-OPfp		1	4	120
	Fmoc-Cys(Trt)-OH	TBTU	1	4	120
10	Fmoc-Asp(OtBu)-OH	TBTU	1	4	90
	Fmoc-Asp(OtBu)-OH	PyBoP	1	4	60
11	Fmoc-Leu-OH	TBTU	1	4	90
12	Fmoc-Gly-OH	TBTU	1	4	90
13	Fmoc-Thr(tBu)-OH	TBTU	1	4	120
	Fmoc-Thr(tBu)-OH	PyBop	1	4	60
	Fmoc-Thr(tBu)-OH	PyBop	1	4	90
14	Fmoc-Phe-OH	TBTU	1	4	90
15	Fmoc-Gly-OH	TBTU	1	4	90

16	Fmoc-Glu(OtBu)-OH	TBTU	1	4	90
17	Fmoc-His(Trt)-OH	TBTU	1	4	120
18	Fmoc-Cys(Trt)-OPfp		1	4	120
	Fmoc-Cys(Trt)-OH	HATU	1	4	120
19	Fmoc-Ile-OH	TBTU	1	4	90
	Fmoc-Ile-OH	PyBop	1	4	90
20	Fmoc-Cys(Trt)-OPfp		1	4	120
	Fmoc-Cys(Trt)-OH	TBTU	1	4	120
21	Fmoc-Gln(Trt)-OH	TBTU	1	4	90
22	Fmoc-Gly-OH	TBTU	1	4	90
	Fmoc-Gly-OH	PyBop	1	4	90
23	Fmoc-Asp(OtBu)-OH	TBTU	1	4	90
24	Fmoc-Val-OH	TBTU	1	4	120
	Fmoc-Val-OH	PyBop	1	4	120
25	Fmoc-Cys(Trt)-OPfp		1	4	120
	Fmoc-Cys(Trt)-OH	TBTU	3	4	120
	Fmoc-Cys(Trt)-OH	PyBop	2	4	90
26	Fmoc-Arg(Pbf)-OH	TBTU	1	4	90
	Fmoc-Arg(Pbf)-OH	PyBop	1	4	60
27	Fmoc-Gly-OH	TBTU	1	4	90
28	Fmoc-Gln(Trt)-OH	TBTU	1	4	90
	Fmoc-Gln(Trt)-OH	PyBop	1	4	90
29	Fmoc-Gly-OH	TBTU	1	4	90
30	Fmoc-His(Trt)-OH	TBTU	1	4	120
31	Fmoc-Cys(Trt)-OPfp		1	4	120
	Fmoc-Cys(Trt)-OH	TBTU	1	4	120
32	Fmoc-Asp(OtBu)-OH	TBTU	1	4	90
33	Fmoc-Ser(tBu)-OH	TBTU	1	4	90

References

1. Cohen, S. (1962) *J. Biol. Chem.*, **237,** 1555.
2. Carpenter, G. and Cohen, S. (1979) Epidermal growth factor. *Annu Rev Biochem*, **48,** 193-216.
3. Walker, P., Weichsel, M.E., Jr., Hoath, S.B., Poland, R.E. and Fisher, D.A. (1981) Effect of thyroxine, testosterone, and corticosterone on nerve growth factor (NGF) and epidermal growth factor (EGF) concentrations in adult female mouse submaxillary gland: dissociation of NGF and EGF responses. *Endocrinology*, **109,** 582-587.
4. Pascall, J.C., Saunders, J., Blakeley, D.M., Laurie, M.S. and Brown, K.D. (1989) Tissue-specific effects of castration and ovariectomy on murine epidermal growth factor and its mRNA. *J Endocrinol*, **121,** 501-506.
5. Valcarce, C., Bjork, I. and 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.
6. 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.
7. Sampoli Benitez, B.A. and Komives, E.A. (2000) Disulfide bond plasticity in epidermal growth factor. *Proteins*, **40,** 168-174.
8. Wells, A. (1999) EGF receptor. *Int J Biochem Cell Biol*, **31,** 637-643.
9. Jorissen, R.N., Walker, F., Pouliot, N., Garrett, T.P., Ward, C.W. and Burgess, A.W. (2003) Epidermal growth factor receptor: mechanisms of activation and signalling. *Exp Cell Res*, **284,** 31-53.
10. Lehto, V.P. (2001) EGF receptor: which way to go? *FEBS Lett*, **491,** 1-3.
11. Lu, H.S., Chai, J.J., Li, M., Huang, B.R., He, C.H. and Bi, R.C. (2001) Crystal structure of human epidermal growth factor and its dimerization. *J Biol Chem*, **276,** 34913-34917.

12. Hognason, T., Chatterjee, S., Vartanian, T., Ratan, R.R., Ernewein, K.M. and Habib, A.A. (2001) Epidermal growth factor receptor induced apoptosis: potentiation by inhibition of Ras signaling. *FEBS Lett*, **491,** 9-15.

13. Olszewska-Pazdrak, B., Ives, K.L., Park, J., Townsend, C.M., Jr. and Hellmich, M.R. (2003) Epidermal growth factor potentiates CCK2/gastrin receptor-mediated Ca2+ release by activation of mitogen-activated protein kinases. *J Biol Chem*.

14. Normanno, N., Bianco, C., De Luca, A. and Salomon, D.S. (2001) The role of EGF-related peptides in tumor growth. *Front Biosci*, **6,** D685-707.

15. Osborne, C.K. and Schiff, R. (2003) Growth factor receptor cross-talk with estrogen receptor as a mechanism for tamoxifen resistance in breast cancer. *Breast*, **12,** 362-367.

16. Kim, H.G., Kassis, J., Souto, J.C., Turner, T. and Wells, A. (1999) EGF receptor signaling in prostate morphogenesis and tumorigenesis. *Histol Histopathol*, **14,** 1175-1182.

17. Giannini, G., Ambrosini, M.I., Di Marcotullio, L., Cerignoli, F., Zani, M., MacKay, A.R., Screpanti, I., Frati, L. and Gulino, A. (2003) EGF- and cell-cycle-regulated STAG1/PMEPA1/ERG1.2 belongs to a conserved gene family and is overexpressed and amplified in breast and ovarian cancer. *Mol Carcinog*, **38,** 188-200.

18. Yarden, Y. (2001) The EGFR family and its ligands in human cancer. signalling mechanisms and therapeutic opportunities. *Eur J Cancer*, **37 Suppl 4,** S3-8.

19. Baselga, J., Pfister, D., Cooper, M.R., Cohen, R., Burtness, B., Bos, M., D'Andrea, G., Seidman, A., Norton, L., Gunnett, K. *et al.* (2000) Phase I studies of anti-epidermal growth factor receptor chimeric antibody C225 alone and in combination with cisplatin. *J Clin Oncol*, **18,** 904-914.

20. Mendelsohn, J. and Baselga, J. (2000) The EGF receptor family as targets for cancer therapy. *Oncogene*, **19,** 6550-6565.

21. Campbell, I.D. and Bork, P. (1993) Epidermal growth factor-like modules. *Curr Opin Struct Biol*, **3,** 385-392.

22. Artavanis-Tsakonas, S., Matsuno, K. and Fortini, M.E. (1995) Notch signaling. *Science*, **268,** 225-232.

23. Kleinman, H.K., Graf, J., Iwamoto, Y., Kitten, G.T., Ogle, R.C., Sasaki, M., Yamada, Y., Martin, G.R. and Luckenbill-Edds, L. (1987) Role of basement membranes in cell differentiation. *Ann N Y Acad Sci*, **513,** 134-145.

24. Lee, B., Godfrey, M., Vitale, E., Hori, H., Mattei, M.G., Sarfarazi, M., Tsipouras, P., Ramirez, F. and Hollister, D.W. (1991) Linkage of Marfan syndrome and a phenotypically related disorder to two different fibrillin genes. *Nature*, **352,** 330-334.

25. Engel, J. (1989) EGF-like domains in extracellular matrix proteins: localized signals for growth and differentiation. *FEBS Lett*, **251,** 1-7.

26. Carpenter, G. (1993) EGF: new tricks for an old growth factor. *Curr Opin Cell Biol*, **5,** 261-264.

27. Swindle, C.S., Tran, K.T., Johnson, T.D., Banerjee, P., Mayes, A.M., Griffith, L. and Wells, A. (2001) Epidermal growth factor (EGF)-like repeats of human tenascin-C as ligands for EGF receptor. *J Cell Biol*, **154,** 459-468.

28. Kurniawan, N.D., Aliabadizadeh, K., Brereton, I.M., Kroon, P.A. and Smith, R. (2001) NMR structure and backbone dynamics of a concatemer of epidermal growth factor homology modules of the human low-density lipoprotein receptor. *J Mol Biol*, **311,** 341-356.

29. Stenflo, J. (1991) Structure-function relationships of epidermal growth factor modules in vitamin K-dependent clotting factors. *Blood*, **78,** 1637-1651.

30. Stenflo, J., Stenberg, Y. and Muranyi, A. (2000) Calcium-binding EGF-like modules in coagulation proteinases: function of the calcium ion in module interactions. *Biochim Biophys Acta*, **1477,** 51-63.

31. Haines, N. and Irvine, K.D. (2003) Glycosylation regulates Notch signalling. *Nat Rev Mol Cell Biol*, **4,** 786-797.

32. Tolkatchev, D. and Ni, F. (1998) Calcium binding properties of an epidermal growth factor-like domain from human thrombomodulin. *Biochemistry*, **37,** 9091-9100.

33. Persson, E., Selander, M., Linse, S., Drakenberg, T., Ohlin, A.K. and Stenflo, J. (1989) Calcium binding to the isolated beta-hydroxyaspartic acid-containing

epidermal growth factor-like domain of bovine factor X. *J Biol Chem*, **264,** 16897-16904.

34. Rees, D.J., Jones, I.M., Handford, P.A., Walter, S.J., Esnouf, M.P., Smith, K.J. and Brownlee, G.G. (1988) The role of beta-hydroxyaspartate and adjacent carboxylate residues in the first EGF domain of human factor IX. *Embo J*, **7,** 2053-2061.

35. Handford, P.A., Mayhew, M., Baron, M., Winship, P.R., Campbell, I.D. and Brownlee, G.G. (1991) Key residues involved in calcium-binding motifs in EGF-like domains. *Nature*, **351,** 164-167.

36. Artavanis-Tsakonas, S., Rand, M.D. and Lake, R.J. (1999) Notch signaling: cell fate control and signal integration in development. *Science*, **284,** 770-776.

37. Blasi, F., Vassalli, J.D. and Dano, K. (1987) Urokinase-type plasminogen activator: proenzyme, receptor, and inhibitors. *J Cell Biol*, **104,** 801-804.

38. Appella, E., Weber, I.T. and Blasi, F. (1988) Structure and function of epidermal growth factor-like regions in proteins. *FEBS Lett*, **231,** 1-4.

39. Hommel, U., Harvey, T.S., Driscoll, P.C. and 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.

40. Cooke, R.M., Wilkinson, A.J., Baron, M., Pastore, A., Tappin, M.J., Campbell, I.D., Gregory, H. and Sheard, B. (1987) The solution structure of human epidermal growth factor. *Nature*, **327,** 339-341.

41. Erickson, H.P. (1993) Tenascin-C, tenascin-R and tenascin-X: a family of talented proteins in search of functions. *Curr Opin Cell Biol*, **5,** 869-876.

42. Clark, R.A., Erickson, H.P. and Springer, T.A. (1997) Tenascin supports lymphocyte rolling. *J Cell Biol*, **137,** 755-765.

43. Tsunoda, T., Inada, H., Kalembeyi, I., Imanaka-Yoshida, K., Sakakibara, M., Okada, R., Katsuta, K., Sakakura, T., Majima, Y. and Yoshida, T. (2003) Involvement of large tenascin-C splice variants in breast cancer progression. *Am J Pathol*, **162,** 1857-1867.

44. Leins, A., Riva, P., Lindstedt, R., Davidoff, M.S., Mehraein, P. and Weis, S. (2003) Expression of tenascin-C in various human brain tumors and its relevance for survival in patients with astrocytoma. *Cancer*, **98,** 2430-2439.

45. Atula, T., Hedstrom, J., Finne, P., Leivo, I., Markkanen-Leppanen, M. and Haglund, C. (2003) Tenascin-C expression and its prognostic significance in oral and pharyngeal squamous cell carcinoma. *Anticancer Res*, **23,** 3051-3056.

46. Chiquet-Ehrismann, R. and Chiquet, M. (2003) Tenascins: regulation and putative functions during pathological stress. *J Pathol*, **200,** 488-499.

47. Cantor, C.R. and Schimmel, P.R. (1980) *The conformation of biological macromolecules*. W. H. Freeman, San Francisco.

48. Petersen, M.T., Jonson, P.H. and Petersen, S.B. (1999) Amino acid neighbours and detailed conformational analysis of cysteines in proteins. *Protein Eng*, **12,** 535-548.

49. Pain, R.H. (2000) *Mechanisms of protein folding*. 2nd ed. Oxford University Press, Oxford ; New York.

50. Creighton, T.E. (1997) Protein folding coupled to disulphide bond formation. *Biol Chem*, **378,** 731-744.

51. Goto, Y. and Hamaguchi, K. (1981) Formation of the intrachain disulfide bond in the constant fragment of the immunoglobulin light chain. *J Mol Biol*, **146,** 321-340.

52. Vanhove, M., Guillaume, G., Ledent, P., Richards, J.H., Pain, R.H. and Frere, J.M. (1997) Kinetic and thermodynamic consequences of the removal of the Cys-77-Cys-123 disulphide bond for the folding of TEM-1 beta-lactamase. *Biochem J*, **321 (Pt 2),** 413-417.

53. Raina, S. and Missiakas, D. (1997) Making and breaking disulfide bonds. *Annu Rev Microbiol*, **51,** 179-202.

54. Kemmink, J., Darby, N.J., Dijkstra, K., Nilges, M. and Creighton, T.E. (1997) The folding catalyst protein disulfide isomerase is constructed of active and inactive thioredoxin modules. *Curr Biol*, **7,** 239-245.

55. Darby, N.J., Kemmink, J. and Creighton, T.E. (1996) Identifying and characterizing a structural domain of protein disulfide isomerase. *Biochemistry*, **35,** 10517-10528.

56. Holmgren, A. (1995) Thioredoxin structure and mechanism: conformational changes on oxidation of the active-site sulfhydryls to a disulfide. *Structure*, **3,** 239-243.

57. Martin, J.L. (1995) Thioredoxin--a fold for all reasons. *Structure*, **3,** 245-250.

58. Kobayashi, T. and Ito, K. (1999) Respiratory chain strongly oxidizes the CXXC motif of DsbB in the Escherichia coli disulfide bond formation pathway. *Embo J*, **18,** 1192-1198.

59. Kishigami, S. and Ito, K. (1996) Roles of cysteine residues of DsbB in its activity to reoxidize DsbA, the protein disulphide bond catalyst of Escherichia coli. *Genes Cells*, **1,** 201-208.

60. Bardwell, J.C., Lee, J.O., Jander, G., Martin, N., Belin, D. and Beckwith, J. (1993) A pathway for disulfide bond formation in vivo. *Proc Natl Acad Sci U S A*, **90,** 1038-1042.

61. Bader, M., Muse, W., Ballou, D.P., Gassner, C. and Bardwell, J.C. (1999) Oxidative protein folding is driven by the electron transport system. *Cell*, **98,** 217-227.

62. Missiakas, D., Georgopoulos, C. and Raina, S. (1994) The Escherichia coli dsbC (xprA) gene encodes a periplasmic protein involved in disulfide bond formation. *Embo J*, **13,** 2013-2020.

63. Bulaj, G., Kortemme, T. and Goldenberg, D.P. (1998) Ionization-reactivity relationships for cysteine thiols in polypeptides. *Biochemistry*, **37,** 8965-8972.

64. Welker, E., Narayan, M., Wedemeyer, W.J. and Scheraga, H.A. (2001) Structural determinants of oxidative folding in proteins. *Proc Natl Acad Sci U S A*, **98,** 2312-2316.

65. Karplus, M. (1997) The Levinthal paradox: yesterday and today. *Fold Des*, **2,** S69-75.

66. Onuchic, J.N., Wolynes, P.G., Luthey-Schulten, Z. and Socci, N.D. (1995) Toward an outline of the topography of a realistic protein-folding funnel. *Proc Natl Acad Sci U S A*, **92,** 3626-3630.

67. Bryngelson, J.D., Onuchic, J.N., Socci, N.D. and Wolynes, P.G. (1995) Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins*, **21,** 167-195.

68. Wolynes, P.G., Onuchic, J.N. and Thirumalai, D. (1995) Navigating the folding routes. *Science*, **267,** 1619-1620.

69. Dobson, C.M. (2003) Protein folding and misfolding. *Nature*, **426,** 884-890.

70. Dinner, A.R., Sali, A., Smith, L.J., Dobson, C.M. and Karplus, M. (2000) Understanding protein folding via free-energy surfaces from theory and experiment. *Trends Biochem Sci*, **25,** 331-339.

71. Dobson, C.M., Sali, A. and Karplus, M. (1998) Protein Folding: a perspective from theory and experiments. *Angew. Chem. Int. Ed.*, 868.

72. Kaiser, E., Colescott, R.L., Bossinger, C.D. and Cooke, P.I. (1970) Color test for detection of free terminal amino groups in the solid-phase synthesis of peptides. *Anal. Biochem.*, **34,** 565.

73. Taylor, J.A., Walsh, K.A. and Johnson, R.S. (1996) Sherpa: a Macintosh-based expert system for the interpretation of electrospray ionization LC/MS and MS/MS data from protein digests. *Rapid Commun Mass Spectrom*, **10,** 679-687.

74. Zanuttin, F., Guarnaccia, C., Pintar, A., Pongor, S. Folding of epidermal growth factor-like repeats from human tenascin studied through a sequence frame-shift approach. *European Journal of Biochemistry*. 2004 Nov; 271 (21): 4229-40.

75. Jansens, A., van Duijn, E. and Braakman, I. (2002) Coordinated nonvectorial folding in a newly synthesized multidomain protein. *Science*, **298,** 2401-2403.

76. Freedman, R.B., Klappa, P. and Ruddock, L.W. (2002) Protein disulfide isomerases exploit synergy between catalytic and specific binding domains. *EMBO Rep*, **3,** 136-140.

77. Freedman, R.B., Klappa, P. and Ruddock, L.W. (2002) Model peptide substrates and ligands in analysis of action of mammalian protein disulfide-isomerase. *Methods Enzymol*, **348,** 342-354.

78. Winter, J., Klappa, P., Freedman, R.B., Lilie, H. and Rudolph, R. (2002) Catalytic activity and chaperone function of human protein-disulfide isomerase are required for the efficient refolding of proinsulin. *J Biol Chem*, **277,** 310-317.

79. Pirneskoski, A., Klappa, P., Lobell, M., Williamson, R.A., Byrne, L., Alanen, H.I., Salo, K.E., Kivirikko, K.I., Freedman, R.B. and 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.

80. Blond-Elguindi, S., Cwirla, S.E., Dower, W.J., Lipshutz, R.J., Sprang, S.R., Sambrook, J.F. and Gething, M.J. (1993) Affinity panning of a library of peptides displayed on bacteriophages reveals the binding specificity of BiP. *Cell*, **75,** 717-728.

81. Kowalski, J.M., Parekh, R.N., Mao, J. and Wittrup, K.D. (1998) Protein folding stability can determine the efficiency of escape from endoplasmic reticulum quality control. *J Biol Chem*, **273,** 19453-19458.

82. Merrifield. (1963) *j.Am. Chem. Soc.*, **85,** 2149.

83. Becker, H., Lucas, H.W., Maul, J., Pillai, V.N.R., Anzinger, H. and Mutter, M. (1982) *Makromol. Chem. Rapid Commun.*, **3,** 217.

84. Hellermann, H., Lucas, H.W., Maul, J., Pillai, V.N.R. and Mutter, M. (1983) *Makromol. Chem. Rapid Commun.*, **184,** 2603.

85. O'Ferral, R.A.M. (1970) *J. Chem. Soc. (B).*

86. Jones, D.A., Mikulec, R.A. and Mazur, R.H. (1973) *J. Org. Chem.*, **38,** 2865.

87. Lloyd-Williams, P., Albericio, F. and Giralt, E. (1997) In C. W. Rees, C., FRS (ed.), *Chemical approaches to the synthesis of peptides and proteins*, London.

88. Sheehan, J.C. and Hess, G.P. (1955) *j.Am. Chem. Soc.*, **77,** 1067.

89. Rebek, J. and Fettler, D. (1974) *j.Am. Chem. Soc.*, 1606.

90. DeTar, D.F. and Silverstein, R. (1966) *j.Am. Chem. Soc.*, **88,** 1020.

91. Merrifield, R.B., Gisin, B.F. and Bach, A.N. (1977) *J. Org. Chem.*, **42,** 1291.

92. Paul, R. and Kende, A.S. (1964) *J. Am. Chem. Soc.*, 4162.

93. Coste, J., Le-Nguyen, D. and Castro, B. (1990) *Tetrahedron Lett.*, **31,** 205.

94. Carpino, L.A., Imazumi, H., El-Faham, A., Ferrer, F.J., Zhang, C., Lee, Y., Foxman, B.M., Henklein, P., Hanay, C., Mugge, C. *et al.* (2002) The

uronium/guanidinium Peptide coupling reagents: finally the true uronium salts. *Angew Chem Int Ed Engl*, **41,** 441-445.

95. Han, Y., Albericio, F. and Barany, G. (1997) Occurrence and Minimization of Cysteine Racemization during Stepwise Solid-Phase Peptide Synthesis(1)(,)(2). *J Org Chem*, **62,** 4307-4312.

96. Hoffmann, E.d., Charette, J.J. and Stroobant, V. (1996) *Mass spectrometry : principles and applications*. Wiley ;
Masson, Chichester ; New York Paris.

97. Yamashita, M. and Fenn, J.B. (1984) Electrospray ion source. Another variation on the free-jet theme. *J. Phys. Chem.*, **88,** 4451-4459.

98. Whitehouse, C.M., Dreyer, R.N., Yamashita, M. and Fenn, J.B. (1985) Electrospray interface for liquid chromatographs and mass spectrometers. *Anal. Chem.*, **57,** 675-679.

99. Zhang, Z. and Marshall, A.G. (1998) A universal algorithm for fast and automated charge state deconvolution of electrospray mass-to-charge ratio spectra. *J Am Soc Mass Spectrom*, **9,** 225-233.

100. Mann, M., Meng, C.K. and Fenn, J.B. (1989) Interpreting mass spectra of multiply charged ions. *Anal. Chem.*, **61,** 1702-1708.

101. Abian, J. (1999) The coupling of gas and liquid chromatography with mass spectrometry. *J Am Soc Mass Spectrom*, **34,** 157-168.

102. Last, A.M. and Robinson, C.V. (1999) Protein folding and interactions revealed by mass spectrometry. *Curr Opin Chem Biol*, **3,** 564-570.

103. Siuzdak, G. (1994) The emergence of mass spectrometry in biochemical research. *Proc Natl Acad Sci U S A*, **91,** 11290-11297.

104. Pramanik, B.N., Bartner, P.L., Mirza, U.A., Liu, Y.H. and Ganguly, A.K. (1998) Electrospray ionization mass spectrometry for the study of non-covalent complexes: an emerging technology. *J Mass Spectrom*, **33,** 911-920.

105. Fasman, G.D. (1996) *Circular dichroism and the conformational analysis of biomolecules*. Plenum Press, New York.

106. Cantor, C.R. and Schimmel, P.R. (1980) *Techniques for the study of biological structure and function*. W. H. Freeman, San Francisco.

107. Perczel, A., Park, K. and Fasman, G.D. (1992) Analysis of the circular dichroism spectrum of proteins using the convex constraint algorithm: a practical guide. *Anal Biochem*, **203,** 83-93.

108. Sreerama, N. and 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.

109. Sreerama, N. and Woody, R.W. (1993) A self-consistent method for the analysis of protein secondary structure from circular dichroism. *Anal Biochem*, **209,** 32-44.

110. Andrade, M.A., Chacon, P., Merelo, J.J. and 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.