Aromatic cluster sensor of protein folding: near-UV ECD bands assigned to fold compactness

Viktor Farkas[a], Imre Jákli[a], Gábor K. Tóth[b] and András Perczel*[a,c]

[a] MTA-ELTE Protein Modelling Research Group, Eötvös Loránd University, Pázmány P. sétány 1/A, Budapest, H-1117 Hungary

[b] Department of Medical Chemistry, University of Szeged, Dóm tér 8, Szeged, H-6720 Hungary

[c] Laboratory of Structural Chemistry and Biology, Eötvös Loránd University, Pázmány P. sétány 1/A, Budapest, H-1117 Hungary

Abstract

Both far- and near-UV ECD spectra have bands sensitive to thermal unfolding of Trp and Tyr residues containing proteins. Beside spectral changes at 222 nm reporting secondary structural variations (far-UV range), L_b bands (near-UV range) are applicable as 3D-fold sensors of protein's core structure. In this study we show that both L_b(Tyr) and L_b(Trp) ECD bands could be used as **sensors of fold compactness.** ECD is a relative method and thus requires NMR referencing and cross-validation, also provided here. The ensemble of 204 ECD spectra of Trp-cage miniproteins is analysed as training set for "calibrating" Trp↔Tyr folded systems of known NMR structure. While in the far-UV ECD spectra changes are linear as function of the temperature, near-UV ECD data indicate a non-linear and thus, cooperative unfolding mechanism of these proteins. Ensemble of ECD spectra deconvoluted gives both conformational weights and insight to protein folding↔unfolding mechanism. We found that the L_b^{293} band is reporting on the 3D-structure compactness. In addition, the pure **near-UV ECD spectrum** of the **unfolded state** is described here for the first time. Thus, ECD folding information now validated can be applied with confidence in a large thermal window (5≤T≤85 °C) compared to NMR for studying the unfolding of Trp↔Tyr residue pairs. In conclusion, folding propensities of important proteins (RNA polymerase II, ubiquitin protein ligase, tryptase-inhibitor *etc.*) can now be analysed with higher confidence.

Abstract in Hungarian / Magyar nyelvű összefoglaló

Az aromás oldalláncot, főként a Trp és Tyr-t – tartalmazó fehérjék távoli- és közeli-UV tartományba eső elektronikus cirkuláris dikroizmus spektruma (ECD) fontos információt ad a makromolekulák le- és feltekeredésének mechanizmusáról. A 222 nm hullámhosszon mért és elterjedten használt ECD jelintenzitás-változás mellett, a Trp és Tyr aminosavak oldalláncainak L_a és L_b ECD sávjai remekül használhatok a fehérje-térszerkezet tömörségének jellemzésére. Mivel az ECD mérésekből levont következtetések viszonylagosak, ezért a kapott eredményeket NMR adatokkal hitelesítettük a Trp-kalitka minifehérje 12 elemű családjában. Alkalmas NMR adatokhoz kalibráltuk a 204 hőmérsékletfüggő ECD spektrum-együttes analízisének eredményeit. Míg a másodlagos térszerkezetre jellemző távoli-UV ECD sávok intenzitása a hőmérséklet emelkedésével lineárisan változik, addig az eltemetett Trp, Tyr oldalláncok ECD sávjainak intenzitásváltozása nem-líneáris profilt mutat, utalva ezzel a téralkat letekeredésének kooperatív jellegére. Az általunk korábban kifejlesztett görbefelbontó program (CCA+) eredményei rámutattak arra, hogy a Trp L_b^{293} sávja alkalmas a (mini)fehérjék harmadlagos szerkezeti tömörségének jellemzésére. Továbbá sikerült elsőként jellemeznünk a fehérjék denaturált téralkatára

jellemző közeli-UV ECD spektrumot. Mostani eredményeink fényében kijelenthetjük, hogy a közeli-UV ECD spektrum alkalmas a Trp↔Tyr aminosav-párt tartalmazó fehérjék térszerkezet-vizsgálatára, mely módszer megbízhatóságát az alkalmazott széles hőmérséklettartomány (5°C≤ T≤ 85°C) tovább fokozza. Az újonnan fejlesztett és hitelesített módszerünk olyan fontos fehérjék hatékony és gyors térszerkezet-vizsgálatát teszi lehetővé, mint a közismert RNS polimeráz II, ubikvitin-ligáz, vagy a triptáz-inhibitor.

Keywords: far- and near-UV ECD spectra, 3D-fold sensor, thermal unfolding, Trp-Tyr interacting residues, folding intermediate,

Abbreviations used:

$[\varTheta]_{MR}$: Mean Residue Molar Ellipticity, $f(T)=[\varTheta(\lambda)]^{T,model}$: temperature and model dependent ECD spectrum, CCA+: Convex Constraint Analysis deconvolution method, **F**-state: folded state, $[F]$: conformational weight of **F**-state, **I**-sate: intermediate/transition state, $[I]$: conformational weight of **I**-state, **U**-state: ensemble of the unfolded state, $[U]$: conformational weight of **U**-state, $X_{F(cage)}$: NMR determined $[F]$ (by summing ^{1}H secondary chemical shifts of selected core residues), $X_{F(helix)}$: NMR determined $[F]$ (by summing ^{1}H secondary chemical shifts of selected residues in helix), $X_F^{average}$ (NMR)= ½$[X_{F(cage)} +X_{F(helix)}]$, $X_{F(CCA\_farECD)}$: $[F]$ determined by deconvoluting ensemble of far-UV ECD spectra, $X_{F(CCA\_nearECD)}$: $[F]$ determined by deconvoluting ensemble of near-UV ECD spectra, $X_F^{average}$(ECD)= ½$[X_{F(CCA\_farECD)} +X_{F(CCA\_nearECD)}]$, T: temperature (°C), T_{opt}: T (°C) with the highest 3D-fold compactness, I_{max}: T (°C) where the relative contribution of **I**-state is the maximum,

Introduction

The biologically relevant and functional form of most of the proteins is the folded, structured state. Unlike intrinsically disordered proteins (IDPs),[1] most globular proteins have a single, well-defined 3D-structure with backbone/sidechain dynamics more prominent at the outer layer of the macromolecule. The inner core is formed by more rigid residues, among which hydrophobic residues such as Phe, Tyr and Trp are more prevalent than charged or hydrophilic amino acids. As a consequence, spatially close Trp and Tyr residues (Trp↔Tyr) are also common.[2] These amino acid pairs behave as chirally perturbed chromophores which absorb UV light with characteristic bands in the near UV spectrum. The fine structure of these bands is a potentially sensitive reporter of the geometrical properties of the interacting phenol and indole rings. However, the assignment of these bands is challenging, since usually more than one Trp↔Tyr pair is present in a protein core.

In order to characterize and assign the near-UV absorption bands of a Trp-Tyr amino acid pair we chose the Trp-cage (TC) miniprotein and its derivatives as a model system, since they have a single Trp-Tyr interacting pair in the hydrophobic core with no other perturbing chromophors. Several TC variants have been structurally characterized by NMR and far-UV ECD spectroscopy in our lab among which the elongated series (E0-E19) showed the most pronounced structural changes from unfolded variant (E0) to highly stable, well-folded protein (E19). The well-characterized gradual structural stabilization and increased thermal resistance among these proteins can be exploited to describe the temperature dependent changes of the near-UV ECD spectrum caused by the Trp-Tyr interaction. The acquired information can facilitate the assignment of the near-UV ECD spectrum of more complex proteins. Furthermore, the Trp-Tyr bands can act as fold-sensors by reporting on the interaction strength among these residues. This behavior will facilitate to study folded, molten globule and unfolded states of various proteins.

We have reported recently a set of TC-foldamers of various length (E0 up to E11) (**Scheme 1**),[3] and have monitored both their helicity and degree of molecular packing also called as 3D-fold compactness. These 12 proteins form the present "test-case" studied here in details. E0 the first element has a highly dynamic and thus rather unfolded structure ($X_{F(cage)}$= 0.25, $X_{F(helix)}$= 0.30) as it has a positively charged Arg residue at its *N*-terminus, mandatory for receptor binding. However, the five residue elongated variant, E5, has a fold structure, $X_{F(cage)}$= 0.92, $X_{F(helix)}$= 0.83: both values close to 1.[3,4] Thus, E0-E11 set makes a complete covering a large conformational space from weakly folded to highly compact protein tertiary structures.

GLP-1	HAEGTFTSDVSS YLEGQAAKEHAWLVKGR
Ex4	HGEGTFTSDLSKQMEEEAVRLFIEWLKNGGPSSGAPPPS
E19	HGEGTFTSDLSKQMEEEAVRLYIQWLKEGGPSSGRPPPS
E11	DLSKQMEEEAVRLYIQWLKEGGPSSGRPPPS
E10	LSKQMEEEAVRLYIQWLKEGGPSSGRPPPS
E9-E6	
E5	EEEAVRLYIQWLKEGGPSSGRPPPS
E4-E1	
E0	RLYIQWLKEGGPSSGRPPPS

Scheme 1. *Primary sequence of GLP-1, Ex4, E19 and analogues, E0-E11, forming the training set: E0-E4 (red, magenta and pink) are poorly folded (more dynamic) while E5-E19 (blue) have more compact 3D-fold.*

It is more common to analyze far-UV (n→π, π→π*) than near-UV (L_a, L_b) ECD electronic transitions for rapid structure determination of proteins in solution.[5] **Far-UV ECD** spectroscopy (185-260 nm) requires only ~50-100 μM solution of protein to quickly and straightforwardly determine the structural characteristics of the polypeptide backbone chain. A U-type ECD curve is associated with a "random coil" or highly dynamic protein structure, also established for Trp-cage miniproteins.[6] Although, E0, the shortest in the series, shows U-type ECD characteristics, weak n→π* (~220 nm) embedded in a strong negative band (~205 nm) followed by a small positive $[\Theta]_{MR}$ value below 190 nm, it also reflects some residual folded fraction as concluded also by NMR (**Figure 1**). If a Trp-cage

miniprotein is mostly folded (*e.g.* E4-E11), a C-type ECD curve is recorded: strong n→π* electronic transition (λ=222 nm) and π→π* exciton couplet ($[\Theta]_{MRmax}$ ~190 and $[\Theta]_{MRmin}$ ~206 nm). The latter ECD type is that of an α-helix of reduced intensities. Values of $[\Theta]_{MR}^{222}$ are commonly used as semi-quantitative measures of the α-helix content of a polypeptide and protein (*vide infra)*.
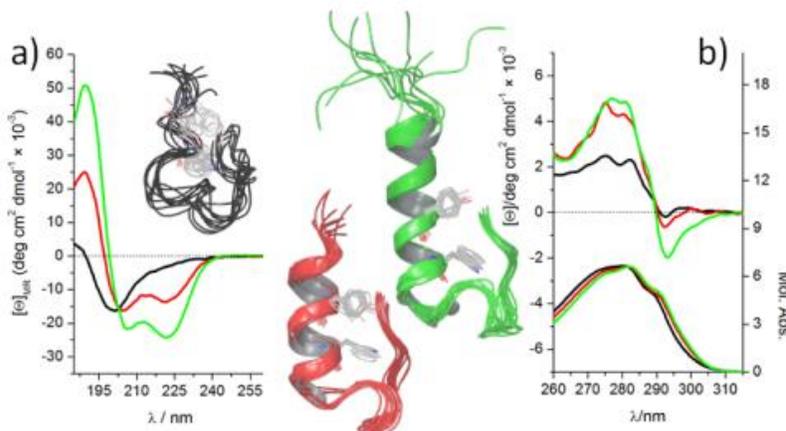
Figure 1. *Ensemble of backbone structures as well as **a)** the far-UV ECD and **b)** the near-UV absorbance (bottom) and ECD (top) spectra of E0 (black), E4 (red) and E11 (green) at 5°C (H₂O: 6.6<pH<7.0). Note that Trp and Tyr sidechains are dynamic in the shorter (e.g. E0, E1), but tightly packed and spaced in Trp-cage models with longer α-helix (E4-E11).*

The **near-UV absorption spectra** of polypeptides and proteins are composed of broad band assemblies, including ^1L_a and ^1L_b electronic transitions originating from aromatic residues, [7] information less elaborated for direct structure elucidation. In Trp↔Tyr nanosystems, the ^1L_b of Tyr, and ^1L_a and ^1L_b of Trp residues are potential sensors of protein folding if the structural markers of the aromatic residues are correlated to spectral properties. However, as these vibration-modulated UV ECD bands highly overlap in the 260-300 nm spectral region their assignment is more complex. As described for the simplest diketopiperazine model,[8] the spatial proximity of Tyr and Trp characteristically shifts the fine structure of the ^1L_b absorption bands of Tyr (~276 nm with a shoulder at ~287 nm) and the ^1L_b band of Trp side-chain (~281 and ~293 nm) (**Figure 2.**), while the ^1L_a band of Trp appears as a very broad band "at the background". The associated near-UV ECD bands provide useful information on both the extent of the packing of the hydrophobic environment of the parent residues and the relative distance /orientation between the Trp and Tyr residues.[8] In line with this, E11 has the largest while E0 the weakest near-UV ECD band intensities, explained as the Trp-Tyr residues are tightly packed against each other and buried from water in E11, unlike in the highly dynamic E0 (**Figure 2.**). These preliminary observations fueled our enthusiasm to elaborate Trp↔Tyr as sensors of 3D fold of proteins, described below.
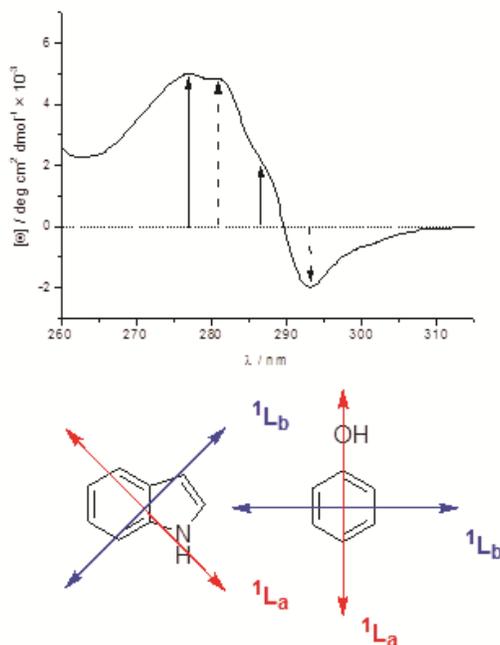
Figure 2. *The near-UV ECD spectrum of E11: arrows indicating the ^1L_b bands of Tyr (276 and 287 nm, solid line) and of Trp (281 and 293 nm, dashed line). ^1L_a (red) and ^1L_b (blue) are the electronic transition moments of indole and phenol rings.*

Our primary goal was to show that a Trp↔Tyr interacting residue pairs are powerful sensors as *i*) they have characteristic near-UV ECD bands signaling their folding degree, *ii*) the arising ECD spectral information can be cross-validated by NMR. The objective of our research was to understand *i*) how secondary structural elements are coupled to each other in a Trp-cage fold, *ii*) how an elongated α-helix can make a TC-fold more heat resistant and *iii*) to what extent these observation are general. In summary, we present here that a near-UV ECD spectrum comprising L_a and L_b vibration modulated electronic transitions is applicable as a sensor to direct 3D-fold characterization of a protein.

Results and Discussion

Band selective analysis of UV-ECD electronic transitions of **Trp↔Tyr** as **3D-fold sensors** of proteins

As for most globular proteins, the compact 3D-fold of TC-miniproteins is gradually lost as T is increased. Heating from 5 to 85°C changes both the n→π* (~222 nm) and π→π* band intensities (~190 nm) and transition occurs *via* a single isodichroic point at ~202 nm (**Figure 3a, SFigure 1**). Thus, the contribution of C-type **far-UV** ECD curve decreases during thermal unfolding while that of the U-type increases. In line with these changes, the folded fraction of the protein, $X_{F(helix)}$, decreases from 1 to 0. The far-UV ECD spectra of the more dynamic E0 – E3 show minor changes as a function of T as these have a U-shape ECD spectra even at low T (5°C). However, even for these weakly folded miniproteins their **near-UV ECD** spectra changes more firmly (**Figure 3b**). For all the 12 miniproteins studied here, E0 – E11, the T-dependent near-UV ECD spectra (270-310 nm) are of interest as both the positive (~276.1±0.7 and ~286.9±0.4 nm of Tyr and ~281.3±0.5 nm of Trp) and the negative vibrational bands (~292.8±0.4 nm) converge to zero at 85°C (**Figure 3b, SFigure 2**).
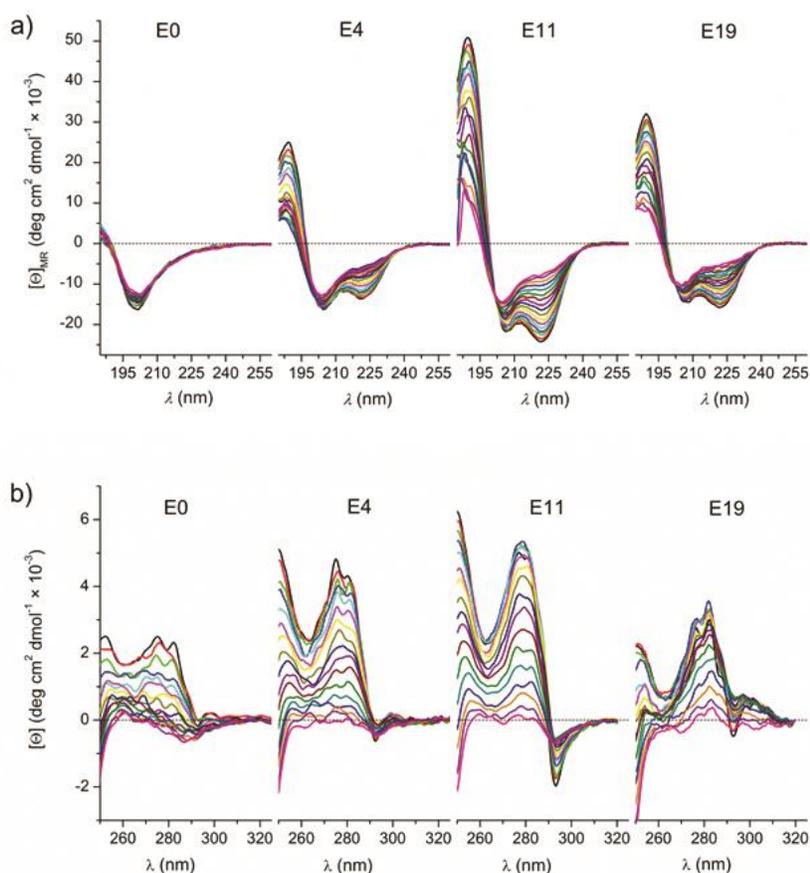
**Figure 3. Figure 3. Following the *temperature dependent helicity and 3D-fold compactness changes of several TC-miniproteins by ECD. a) far-UV and b*) near-UV ECD spectra of selected miniproteins of increasing length (5°C → 85°C, pH: 6.5-7.0 in water).*

For the successful spectroscopic application of Trp↔Tyr as a fold sensor, one has to know which of the electronic transitions carries important and independent (non-redundant) structural information. From the spectroscopist's point of view the first choice would be to monitor the mean residue ellipticity changes at

λ=222 nm in the far-UV ECD spectra (**SFigure3**). At this wavelength, ellipticity changes, d$[\Theta]^{222}$/dT, are primarily determined by shifts in relative orientation of the amide chromophores manifested in variation of n→π* transitions.[9] Intimately linked to 3D-fold compactness, changes of secondary structural elements, mainly α-helices in TC miniproteins are informative.[3]

TC-model [a]		$n\to\pi^*$ ECD $\lambda\sim$222 nm		
		m [b]	b	R^2
fully or partly **unfolded** at 5°C	E0	-3.2	-2855.2	0.217
	E1	5.2	-3595.4	0.467
	E2	11.0	-4912.3	0.714
	E3	50.2	-7947.2	0.939
folded at 5°C	E4	115.2	-13995.5	0.982
	E5	154.5	-19074.1	0.997
	E6	133.5	-18683.7	0.992
	E7	130.1	-17677.6	0.983
	E8	146.8	-19890.1	0.992
	E9	184.5	-22725.6	0.999
	E10	178.3	-21923.1	0.997
	E11	217.8	-25944.8	0.996
folded at 5°C	E19	134.0	-16466.0	0.999
	Ex4	114.0	-13253.1	0.995
	Tc5b	136.1	-13851.3	0.985
	Tc5b_D9E	96.8	-11320.5	0.989

[a] Trp-cage models: E0-E11 plus E19, Ex-4, Tc5b and Tc5b_D9E
[b] steepness value m, dimension: deg.cm^2.dmol^{-1}.°C^{-1}*10^{-3}

Table 1. *Parameters (m, b, R^2) associated with fitting a linear function $[\Theta(T)]^{222,model} = mT + b$ to n→π* ECD band intensities as function of temperature, for 16 TC-miniproteins*

To gain a more quantitative picture, the temperature dependent ellipticity values at λ=222 nm, $[\Theta(T)]^{222,model}$, were plotted against T(°C) and a linear function was fitted for each protein model: E0-E11, E19, Ex-4, Tc5b and Tc5b_D9E. For most TC-miniproteins R^2 >0.939 (**Table 1**). By analyzing m (steepness) and b of the $[\Theta(T)]^{222,model} = mT + b$ linear function quantitative insight is gained on how secondary structure unfolding occurs.
   1) For shorter and thus highly dynamic miniproteins, E0-E3, the steepness of the $f(T)=[\Theta(T)]^{222,E0-E3}$ linear function, m, is small, close to zero: *e.g.* m^{E1}= 5.2 deg.cm^2.dmol^{-1}.C^{-1}*10^{-3}. Thus, in unstructured TC-miniproteins the α-helix content is small (or insignificant) even at low temperature and thus their changes are uninformative.
   2) More importantly, for longer and folded proteins, E4- E11, $f(T) = [\Theta(T)]^{222nm,E4-E11} = mT + b$ function is *quasi* linear with a characteristic steepness m of about 154±22 deg.cm^2.dmol^{-1}.C^{-1}*10^{-3}. Model E5-E10 contains about the same amount of α-helix at lower temperature, which unfolds at about the same rate in all of these proteins as T increases.

Consequently, thermal unfolding of secondary structural elements of these miniproteins is *quasi*-linear and thus their folding/unfolding is non-cooperative. In other words, structural changes may occur at any residue of the constituent α-helices or even simultaneously at multiple sites. However, these folding/unfolding events are independent of each other: no subunit interaction enhances or inhibits micro-exchanges of folding states at any temperature.

Signal intensity changes of the **near-UV** ECD, d$[\Theta]^{Lb-La}$/dT, manifest in form of selected vibrational modes. Unlike E0, miniprotein E4, E11 and E19 are all folded but to a different extent and thus their thermal unfolding is expected to be different. Four characteristic wavelengths were selected to monitor structure unfolding, namely ~277, ~287 nm (of Tyr) and ~281, ~293 nm (of Trp residues). Term d$[\Theta]^\lambda$/dT quantitatively describes such a change, which is the smallest for the most dynamic E0, as d$[\Theta]^\lambda$/dT~0 at

any λ of interest (**SFigure3** and **Figure 4**). However, thermal unfolding of E4, an E0 analogue lengthened by EEAV at its *N*-terminus (**Scheme 1**), is different as it has already a folded TC-structure.[3] Thus, comparing d$[\Theta]^\lambda$/dT terms of poorly folded to folded protein structures, the following is seen. *i*) For more dynamic (weakly folded) miniproteins (*e.g.* E0-E3) at all 4 selected near-UV ECD bands rapid intensity decrease is seen in a quasi-monotonic manner as T increases. *ii*) On the contrary, for compact and highly folded TC-miniproteins (*e.g.* E11), increasing T (5°C<T <30°C) follows first $[\Theta]^{\text{Lb-La}}$ band intensities rise (**Figure 4**). Thus, 3D-fold gets more compact at 30 °C, compared to values at 5°C. Passing beyond this critical point (T_{opt}), the temperature where fold compactness is the highest and $X_{\text{F(cage)}}$ is the closest to 1, 3D-fold starts to degrade and especially at higher T (30°C<T <85°C) becomes disintegrated. Fitting a

Gaussian function, $f(T) = [\Theta(T)]^{\lambda,E4-E11} = a \exp\left(-\frac{(T-T_{opt})^2}{2c^2}\right) + d$, gives T_{opt} at each selected

wavelengths, associated with Tyr and Trp electronic transitions (**Table 2**). Clearly, T_{opt} depends on both the wavelength used to monitor unfolding and on the TC-model. T_{opt} is informative at any λ, but values determined by fitting data to L$_b$(Tyr) λ~287 nm ECD bands are the most realistic one: 24°C$\leq T_{\text{opt}}^{287\text{nm}} \leq$ 32°C (**Figure 5**). Though T_{opt} could be scaled, the relative order of T_{opt} values is already rather informative on 3D-fold compactness.
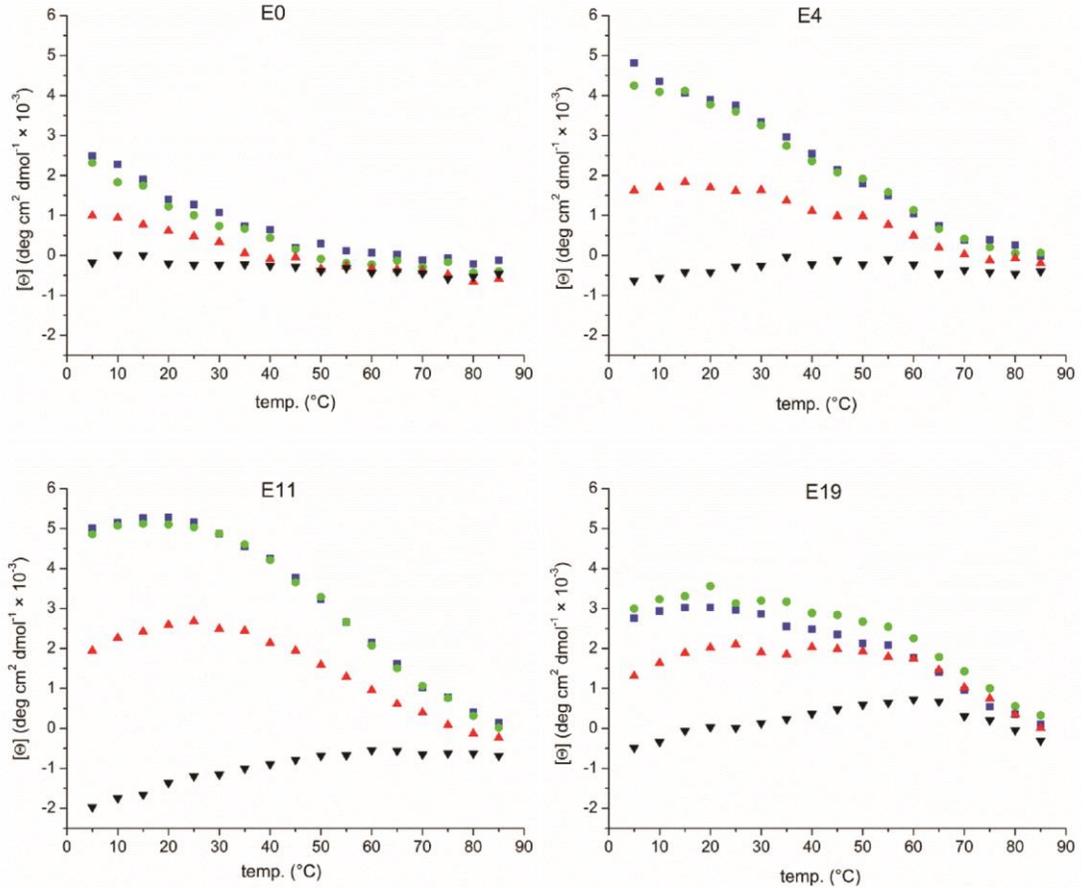
Figure 4. *Measurements of temperature induced unfolding of Trp-cages (E0, E4, E11 and E19) at selected bands associated with Tyr and Trp L$_b$ electronic transitions (near-UV ECD bands). L$_b$ rotational modes of Tyr side-chain at ~277 (blue) and ~287 nm (red), while those of Trp side-chain at ~281 (green) and ~293 nm (black) are indicated.*

	L_b transition [a]	$[\Theta(T)]^{\lambda,modell} = a \exp\left(-\frac{(T-T_{opt})^2}{2c^2}\right) + d$ [b]				
		Tyr (0→+800)	Tyr (0→0)	Trp (0→+850)	Trp (0→0)	common value
		λ= 277 nm	λ= 287 nm	λ= 281 nm	λ= 293 nm	λ= 280 nm
TC-model		T_{opt}	T_{opt}	T_{opt}	T_{opt}	T_{opt}
fully or partly **unfolded** at 5°C	**E0**	-26.1	-11.3	-27.7	-0.1	-14.8
	E1	-51.5	-25.5	- [c]	15.6	-30.5
	E2	- [c]	- [c]	-56.5	8.9	-52.7
	E3	-10.2	6.2	-3.5	28.7	-3.8
folded at 5°C	**E4**	-4.2	15.5	3.1	44.8	1.3
	E5	6.6	24.4	14.2	58.1	13.1
	E6	11.2	32.0	15.2	67.0	15.2
	E7	14.6	30.3	16.8	60.2	17.3
	E8	11.7	31.3	17.0	70.9	16.2
	E9	16.2	30.8	16.6	65.6	17.9
	E10	18.3	26.2	20.0	74.2	18.2
	E11	17.3	24.6	18.1	67.9	18.2
folded at 5°C	**E19**	17.7	34.6	21.8	52.9	21.8
	Tc5b	6.5	20.8	9.8	49.7	10.2
	Tc5b_D9E	-6.3	12.0	0.8	35.8	4.0

[a] [1]L_b electronic transitions at selected vibrational modes: λ ~277, ~287, ~281, ~293 nm plus at the "typical" λ= 280 nm
[b] additional parameters (a, c and d) see *Supporting Information*
[c] If T_{opt} unrealistically shifted for weakly folded Trp-cage models values are not reported here.

Table 2. *Centers (T_{opt}) of Gaussian functions fitted to selected ^1L_b ECD band changes as function of temperature, f(T)=[Θ(T)]$^{\lambda,model}$ of different Trp-cage models: E0-E11 plus E19, Ex-4, Tc5b and Tc5b_D9E*
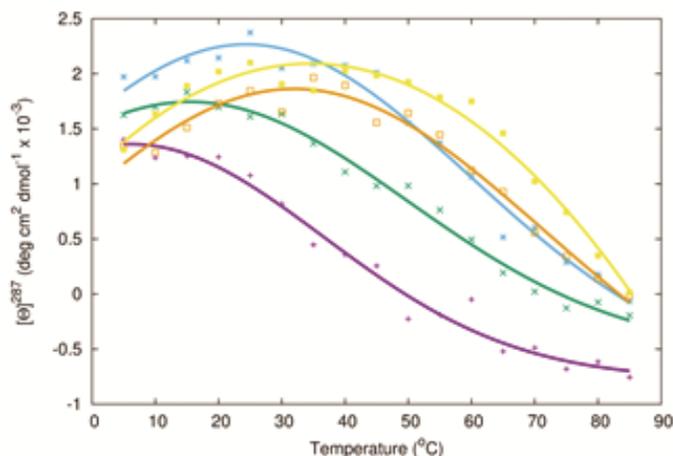
Figure 5. *Tyr (0→0) L_b ECD band (λ=287 nm) changes as function of temperature, [Θ]^{287,model}~f(T), for E3, E4, E5, E6 and E19 TC-models. Note the shift of T_{opt} to higher value indicates higher thermostability of the protein.*

Both far- and near-UV ECD spectra have selected bands sensitive to thermal unfolding. Data acquired at λ=222 nm monitor secondary structure "melting" and signals a **linear** or **non-cooperative** unfolding mechanism, as d[Θ]^{222,model}/dT~ m turns out to be about the same for most folded TC-models: m= 154±22 deg.cm^2.dmol^{-1}.C^{-1}*10^{-3}. On the contrary, especially for the well-folded miniproteins the analysis of L_b bands, signal a **non-linear** or **cooperative** structure unfolding mechanism, with an average $T_{opt}^{287,model}$ = 29±3 °C. Though determining the temperature, T_{opt}, where folding is the most compact, is straightforward, however understanding the molecular background of such a cooperative thermal unfolding triggered by non-linear effects is more of a challenge, *vide infra*.

Part of our sensor validation cross-correlation between the $f(T)=[\Theta(T)]^{\lambda,model}$ functions were made. Novelty in terms of information content of these ECD bands are expressed by cross-correlation coefficients, R^2, determined for all 16 TC-miniproteins at 5 different wavelengths. Data of E11, a typical example of folded TC-miniprotein is discussed below (**Table 3**).

Correlation of L_b(s)	$[\Theta]^{222}$	L_b(Tyr) $\lambda\sim277$ nm	L_b(Trp) $\lambda\sim280$ nm	L_b(Trp) $\lambda\sim281$ nm	L_b(Tyr) $\lambda\sim287$ nm
L_b(Tyr) $\lambda\sim277$ nm	0.956	1.000			
L_b(Trp) $\lambda\sim280$ nm	0.947	0.998	1.000		
L_b(Trp) $\lambda\sim281$ nm	0.948	0.998	1.000	1.000	
L_b(Tyr) $\lambda\sim287$ nm	0.859	0.968	0.973	0.974	1.000
L_b(Trp) $\lambda\sim293$ nm	0.792	*0.615*	*0.597*	*0.597*	*0.450*

Table 3. R^2 *cross-correlation coefficients of* $f(T)=[\Theta(T)]^{\lambda,E11}$ *functions for model* ***E11,*** *a typical folded Trp-cage model.*

1) Unlike the L_b^{293}(Trp) band, other L_b(s) cross-correlate at high significance: $R^2 \geq 0.9$. (**Table 3** and **Figure 4**). Thus, *T*-dependent unfolding information derived from L_b^{277}(Tyr), L_b^{287}(Tyr) and L_b^{281}(Trp) bands is redundant, giving similar type of structural information. Thus, analyzing $f(T)=[\Theta(T)]^{280,model}$ only as the most frequently used aromatic ECD band is adequate.

2) As the shape and sign of the $f(T)= [\Theta(T)]^{293,model}$ is different from the other 3 L_b bands, cross-correlation is less significant, $0.450 \leq R^2 \leq 0.615$, the L_b(Trp) band (vibr. 0-850 cm^{-1}) holds new information.

3) $f(T)=[\Theta(T)]^{222,model}$ correlates significantly with any of the 4 L_b band, as the weakest correlation is $R^2=0.792$ ($[\Theta(T)]^{293,model} \leftrightarrow [\Theta(T)]^{222,model}$).

In conclusion, the temperature dependent $\mathbf{L_b^{293}(Trp)}$ and $\mathbf{L_b^{277}(Tyr)}$ bands give maximum information on the unfolding mode of a Trp-cage miniprotein's core. Using both far-UV and near-UV ECD spectra the use of three wavelengths ($\lambda\sim222$, 277 and 293 nm, **Table 3**) is proposed. Although $[\Theta(T)]^{222}$ correlates with $[\Theta(T)]^{277}$ significantly, $R^2=0.956$, their simultaneous use has the advantage that the former one gives information on "linear", while the latter one on the "non-linear" part of the thermal unfolding (see above).

So far the electronic band selective UV-ECD analysis of Trp↔Tyr sensors as measures of protein fold revealed that *i*) using far-UV spectral data (λ=222 nm) thermal unfolding is non-cooperative while *ii*) data taken from the near-UV spectral range ($\lambda\sim277$, 281, 287 and ~293 nm) suggests a cooperative unfolding mechanism for TC-foldamers. This apparent contradiction is because the former data set (**SFigure 1**) is based on the secondary while the latter one (**SFigure 2**) on tertiary structure changes of proteins. Moving away from the "classical" spectroscopist's approach of analyzing selected and specific bands only, we will carry on by compiling the entire spectral window, of all TC-models with no preselection at all, to reach far-fetched conclusions.

Quantitative analysis of UV-ECD electronic transitions of **Trp↔Tyr** as **3D-fold sensors** of proteins

By analyzing the entire spectral range, $185 \leq \lambda \leq 330$ nm, using a suitable deconvolution method, such as the Convex Constraint Analysis Plus (CCA+) algorithm [10] quantitative information is obtained on

the temperature dependent pure conformational weights: $w^{model}(T)$. Simultaneously, the joint analysis comprising a total 204 (=12*17) far-UV and 204 near-UV ECD spectra gives either two (**Figure 6**) or three (**Figure 7**) pure ECD component curves. This is in line with an **F↔U** (two state, linear transition) (**Figure 6a**) or with a three state, **F↔I↔U** (or cooperative) transition from folded to unfolded state. In general, temperature dependent pure conformational weights are obtained, $w^{model}(T)$, with respect to conformer type **F**, **U** and **I** for each of the proteins. For example, $[F]^{E4}(T)$ is a vector comprising the pure conformational weights of the folded state of E4 as function of temperature (5°C≤ T≤ 85 °C) (**Figure 6b,** red squares). The assignment of the C- and U-type pure far-UV ECD curves to protein secondary structures (**Figure 6**) is based on literature data.[5a,10c] As the 'proof of concept' all the 204 experimental spectra can be reconstructed as the linear combination of the C- and U-type **pure** far-UV ECD spectra (**Figure 6a**, red and black) with the corresponding $[F]^{model}(T)$ and $[U]^{model}(T)$ conformational weights. The same holds for the near-UV-ECD spectrum ensemble. The latter data were cross-checked and reinsured by NMR (ambient pH, 5°C).[4c] Results discussed below are grouped into two sections: the case of unfolded or weakly folded (E0-E3) and those of the folded (E5-E11) TC-proteins.

The quantitative analysis of the most dynamic E0 shows, that *i*) it has no or insignificant amount of secondary structural element (even at lower T), *ii*) but has however a semi-packed molten-globule core structure, in which the Trp↔Tyr sensor has still a native like structure. Increasing temperature gradually unpacks the tertiary structure. (**Figure 6d**). ECD based deconvolution gives for E0 a higher % of folded state, $[F]^{E0}(5°C≤T≤30°C)$~ 0.45 than NMR does at ambient temperature, $X_{F(cage)}$~0.25, (**Figure 3b**).[4c] It is important to note that a native-like hydrophobic core of E0 is still present, even if no secondary structural elements (*e.g.* α-helix) can be assigned (hydrophobic collapse). This has implication on the kinetics of Trp-cage unfolding mechanism *vide infra*.[11] At low T (*e.g.* 5°C) where chemical exchange allows easier conformer determination by NMR, it is possible to compare CCA+ assisted pure conformational weights retrieved from ECD to their NMR counterparts for most TC-miniproteins. NMR derived conformational properties are found to match at a significant level with structural data obtained by deconvoluting both far-UV and near-UV ECD spectra and comparison is numerically close if raw data are averaged (**Table 4**).

Name	NMR			far-UV ECD	near-UV ECD	ECD
	$X_{F(cage)}$ [b]	$X_{F(helix)}$ [c]	$X_F^{average}$ (NMR)	$X_{F(CCA+ \ farECD)}$ [d]	$X_{F(CCA+ \ nearECD)}$ [d]	$X_F^{average}$(ECD)
E0	0.25	0.3	**0.28**	0.01	0.59	**0.30**
E1	0.3	0.35	**0.33**	0.05	0.64	**0.35**
E2	0.39	0.4	**0.40**	0.08	0.62	**0.35**
E3	0.64	0.55	**0.60**	0.22	0.72	**0.47**
E4	0.84	0.85	**0.85**	0.49	0.90	**0.70**
E5	0.92	0.83	**0.88**	0.69	0.97	**0.83**
E6	0.94	0.86	**0.90**	0.71	0.90	**0.81**
E7	0.94	0.89	**0.92**	0.69	0.84	**0.77**
E8	0.94	0.89	**0.92**	0.74	0.80	**0.77**
E9	0.94	0.88	**0.91**	0.90	0.76	**0.83**
E10	0.97	1.00	**0.99**	0.85	0.94	**0.90**
E11	0.92	0.92	**0.92**	1.00	0.98	**0.99**
E19	0.88(27°C)	0.89(27°C)	**0.89**	0.65	0.53	**0,59**

[a] unlike marked, all measurements completed at T= 5°C

[b] folded fraction determined by NMR based on secondary chemical shifts of selected residues of the Trp-cage hydrophobic core[3]

[c] folded fraction determined by NMR based on secondary chemical shifts of selected helical residues of the miniprotein[3]

[d] ECD spectra of the ensemble of Trp-cage miniproteins, E0-E11 were deconvoluted by CCA+ resulting in $X_{F(CCA\_farECD)}$ and $X_{F(CCA\_nearECD)}$) conformational weights of **F**-state. Note that elongation at the *N*-terminus of the polypeptide chain stabilizes the fold, and thus, both $X_{F(cage)}$ and $X_{F(helix)}$ values increases from E0 to E4-E6.

Table 4. *Folded fractions cross-validated determined by NMR ($X_{F(cage)}$, $X_{F(helix)}$) and ECD* [a]

Unfolding mechanism of proteins is straightforward to study by ECD as the method offers a larger thermal window (5°C≤ T≤ 85 °C) than NMR does but gives in return low-resolution structure information.

NMR is capable to provide an atomic resolution description on the most pronounced conformer, however it has severe limitations as the narrower temperature range (*e.g.* 5-45°C), line broadening, exchange and loss of sensitivity makes characterization of minor conformers troublesome.[4c] Folded fractions, X_F, for each of the 12 Trp-cage models were determined by NMR and sorted as $X_{F(cage)}$ and $X_{F(helix)}$ (**Table 4**). The former measures the folded fraction *via* the compactness of the hydrophobic core, while the latter one *via* the tightness of the α-helices. These NMR based fold measures are to cross-validate the conformational weights of **F**-state determined by near-UV and far-UV ECD, $X_{F(CCA\_farECD)}$ and $X_{F(CCA\_nearECD)}$. Correlating $X_{F(cage)}$ and $X_{F(CCA\_nearECD)}$ values for all 12 models gives $R^2(X_{F(cage)} \leftrightarrow X_{F(CCA\_nearECD)})$= 0.75 indicating that the NMR-based high resolution conformational weights on the folded state of the proteins do match with those retrieved from near-UV ECD spectra deconvolution. Higher correlation was established if the folded fraction of TC-miniproteins were determined by comparing their helix content form NMR, $X_{F(helix)}$, and their far-UV ECD data, $X_{F(CCA\_farECD)}$, resulting in $R^2(X_{F(helix)} \leftrightarrow X_{F(CCA\_farECD)})$= 0.91. The resulted in X_F values are from different spectral ranges, $X_{F(CCA\_farECD)}$ and $X_{F(CCA\_nearECD)}$, or from different type of NMR chemical shifts: $X_{F(cage)}$ and $X_{F(helix)}$. These reflect to a somewhat similar conformational property of a protein but derived from a different kind of spectral data. If, however, their averages are compared, $X_F^{average}$ (NMR) and $X_F^{average}$ (ECD), correlation is significant: $R^2(X_F^{average}(NMR) \leftrightarrow X_F^{average}(ECD))$= 0.92. Thus, unquestionably for all 12 TC-foldamers, deconvolution resulted in ECD data that are validated by high-resolution NMR information.
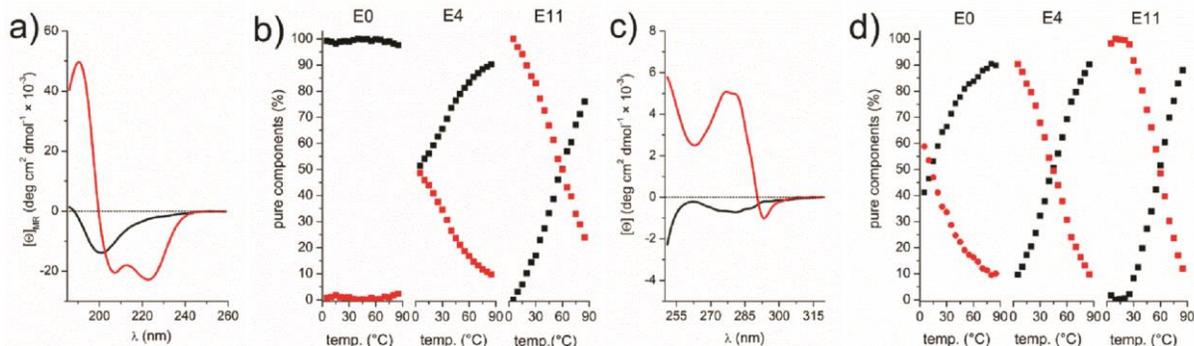
Figure 6. *Selected data from the joint deconvolution of a total of 204 far-UV-ECDs and 204 near-UV-ECDs of 12 miniproteins (E0, E1,…, E10, E11). **a**) The two pure far-UV ECD curves, **b**) their relative temperature dependent conformational weights, **c**) the two pure near-UV ECD curves, **d**) their relative temperature dependent conformational weights, [F]^{model}(T) (red) and [U]^{model}(T) (black) are given for E0, E4 and E11. (ECD-measurement in water (6.5≤pH ≤7.0 ranging 5≤ T ≤85 °C)*

ECD spectra deconvoluted into three pure components are in line with a three-state unfolding model, comprising an **F**-, **I**- and **U**-state. The associated temperature dependent conformational weights are $[F]^{model}(T)$, $[I]^{model}(T)$ and $[U]^{model}(T)$. The NMR based temperature dependent conformer analysis of Tc5b,[4c] the reference of TC-miniproteins,[12] has shown that **I** is an ultrafast on-pathway folding intermediate transiently present during T-induced unfolding.[4c] The present study extends these NMR based unfolding results both in temperature range and number of proteins. E0-E3 miniproteins are omitted from the forthcoming ensemble analysis as the three-state **F**→**I**→**U** unfolding path "starts" with the folded state is underrepresented in these TC-foldamers. Thus, the joint deconvolution with CCA+ was completed for 2×136 UV-ECD spectra (**Figure 7**). The comprehensive analysis shows that **F**- and **I**-states have similar C-type ECD curves but of different n→π* (~222 nm) and π→π* band intensities (~208 and 190 nm, **Figure 7a, b**). Spectral contribution of the **I**-state (**Figure 7a,b,** blue curve) is smaller for the 'just-folded' E4 but more significant for the nicely folded E5-E11 proteins. However, for all proteins, the relative contribution of **I** remains in the range of 10-25% even at the most favorable temperature range (**Figure 7b**). E5 is a typical example containing ~20% of **I**-tate between 30 and 70 °C.
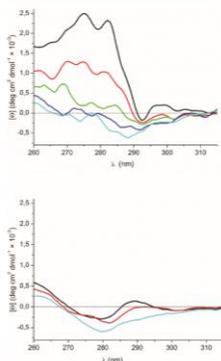
Assignment of the pure **near-UV component** curves (within the 136 near-UV ECD spectra) is less straightforward, as these ECD transitions are less intensive, fewer literature data are available, with spectral properties under continuous revision. The pure near-UV ECD component curve associated with the **I**-state is composed of two positive maxima (~277 and above ~280 nm) and a tiny negative maximum at ~293 nm (**Figure 7c, d** blue) although this component represents a compact-like Trp-cage core structure with a native-like spatial organization of Trp and Tyr residues. This tertiary structure assignment was facilitated by high resolution NMR data (*T*=5°C, *pH*~7) (**Figure 3b**) published earlier.[3] Note that this near-UV pure ECD curve however consists of a weak negative L_b band (~293 nm) only. The second pure near-UV ECD curve is attributed to the compactly folded **F**-state, which has similar spectral features to that of the **I**-state, both in shape and intensities but the latter one has an intensive L_b band at ~293 nm (red / **Figure 7c,d**). One may wonder whether these two rather similar pure near-UV ECD curves of the **I-** and **F**-states stand for two really distinguishable conformers or perhaps these are just spectral deconvolution artifacts. The comprehensive analysis of the present miniprotein family based on their ECD spectral shifts enables us answering this challenge. The 3D-structure analysis and ECD spectra assignment of E11, one of the most tightly folded miniprotein (**Figure 3b**) with known high resolution NMR structure can bring the answer closer (**Figure 1**).[3] The structural differences between E11 and the more weakly folded E4 are subtle but significant. Tightness and molecular packing of the hydrophobic core of E4 and E11 were revealed by NMR (at *T*=5°C) and found as follows: $X^{E4}_{F(cage)}$~0.84 while $X^{E11}_{F(cage)}$~0.92. E11 has a tighter 3D-core structure compared to E4 in which Trp and Tyr residues interact more firmly and thus stronger aromatic 'interaction' enhances the intensity of the L_b electronic transition at 293 nm. Indeed, the conformational weight determined by CCA+ is in line with such a description. In addition, within such a more tightly packed aromatic network fewer or perhaps insignificant amount of waters could be present shifting and enhancing the spectral properties of the L_b (and perhaps L_a) electronic transitions. These factors certainly enhance the intensity of the negative L_b band giving rise to a characteristic near-UV ECD spectrum prototype determined here. In conclusion, the presence of the herein assigned negative L_b band at around 293 nm is probably **a new measure of the tightness and compactness of a Trp↔Tyr protein core**.
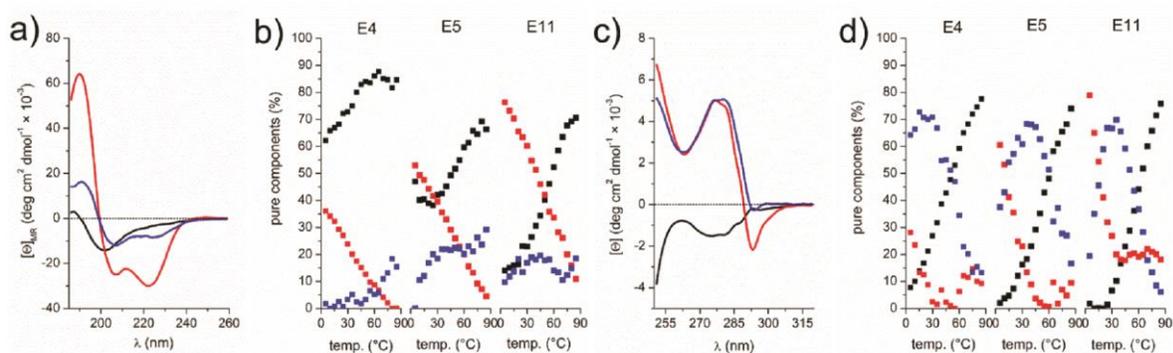
Figure 7. *Selected data from the joint deconvolution of a total of 136 far-UV-ECDs and 136 near-UV-ECD of 8 miniproteins (E4, E5,…, E10, E11). **a**) the three pure **far**-UV ECD curves, **b**) their relative temperature dependent conformational weights **c**) the three pure **near**-UV ECD curves **d**) their relative temperature dependent conformational weights. [F]^model(T) (red), [I]^model(T) (blue) and [U]^model(T) (black) are given for E4, E5 and E11. (ECD-measurement in water (6.5≤ pH ≤7.0 ranging 5≤T ≤85 °C)*

The third and indeed different near-UV pure ECD spectrum obtained from deconvolution is a broader and moderately intense negative band with a maximum at ~280-283 nm (black / **Figure 6c,d** and **7c,d**). Assignment of this pure ECD curve is even more of a challenge solved here based on spectral similarities. The near-UV ECD spectral properties of E0 (**Figure 8a**) and E1, especially at higher *T* (cyan / **Figure 8a**), match with this third pure far-UV ECD curve. Unstructured and thus weakly packed E0 and

13

E1 proteins have a dynamic Trp↔Tyr core (**Figure 1**) with considerable amount of water molecules surrounding and hydrating the aromatic residues. However, the exact amount of unfolding of the latter Trp-cages, as well as their spectral properties remains enigmatic (**Figure 8a**). To conclude the present 'mystery', near-UV ECD spectra of an intrinsically unstructured GGYGGWGG octapeptide contacting both Trp and Tyr residues were measured (water, *pH* ~7). At lower *T,* [Θ] values of this octapeptide are small (0±0.2 deg cm^2 dmol^{-1} 10^{-3}) (5°C, black / **Figure 8b**), L_b electronic transitions are weak and L_a transitions are shifted. As *T* increases, however, spectrum gets more intense (higher negative values), values of [Θ] increase (-0.2±0.2 deg cm^2 dmol^{-1} 10^{-3}) instead of converging to zero. At 25°C (red), the negative band gets broader, while at 85°C (cyan / **Figure 8b**), the octapeptide, most probably fully unfolded, gives an intense negative near-UV ECD curve. The latter spectral features are in good agreement with the deconvolution resulted in third pure near-UV ECD spectrum of Trp-cage miniprotein unfolded. Now confidently assigned, this curve stands for an unfolded hydrophobic core structure comprising a Trp and a Tyr pair of highly dynamic nature, intensively hydrated, termed as the **U**-form.

Figure 8. *Temperature dependent near-UV spectra of **a**) E0 and **b**) GGYGGWGG peptide, (black: 5°C, red: 25°C, green: 45°C, blue: 65°C and cyan: 85°C; in water, at pH: 6.7)*

As the assignment of the three pure near-UV ECD curves are completed, the relative contribution of the pure conformers is to be discussed (**Figure 7d**). At lower *T* (*e.g.* 5 °C), E4 is predominantly folded and thus the relative contribution of the third pure component, **U-form** is small ([U]^{E4}(*T*=5°C)~ 8%) but increases sharply with rising temperature (black / **Figure 7d**). For longer models, E5 - E11, the contribution of the very same **U** form increases as *T* goes up, although in a different manner. The longer the α-helix, the more compact the Trp-cage hydrophobic core is and thus the smaller the relative weight of the unstructured form, [U]^{model}(*T*), is found. For example, at 25°C both E4 and E5 have some unstructured character, [U]^{E4}(*T*=25°C)~ 25%, [U]^{E5}(*T*=25°C) ~15% while the folded E11 has practically zero: [U]^{E11}(*T*=25°C)~ 0-2%. Furthermore, at 5°C contribution of the E4 **F**-state is smaller, [F]^{E4}(*T*=5°C)~30% but as the TC-foldamer gets longer and the overall packing of the Trp-cage gets tighter contribution of **F** gets more and more important: [F]^{E5}(*T*=5°C)~62% and [F]^{E11}(*T*=5°C)~80% (red / **Figure 7d**).

Finally, unlike for **F-** and **U-**states, the temperature dependent weights of the **I**-states, [I]^{model}(*T*), are non-linear (blue / **Figure 7d**) as function of *T*. Contribution of **I** is minor both at lower and higher temperatures but gets more significant at around room *T*. Thus, the relative contribution of conformer **I** **has a maximum**. Furthermore, the temperature preference of **I** having the highest ratio at around 25±15 °C is model dependent. For a short but already folded model, E4, the maximum contribution of **I** is at T~21 °C: [I]^{E4}(*T*=21°C)~70%, thus [I]$_{max}^{E4}$~21°C. As the protein packing gets tighter due to its elongated *N*-terminal α-helix, the temperature where **I** contributes the most is shifted to a higher *T*, namely to 36 or even 44°C: [I]$_{max}^{E5}$~36°C and [I]$_{max}^{E9}$~44°C (**STable 1**). The relative contribution (%) of **I** at its maximum value, **I**$_{max}$, is about the same for all these TC-foldamers, namely 70%: [I]$_{max}^{E4}$~70%, [I]$_{max}^{E5}$~70% and [I]$_{max}^{E9}$~70%. We conclude that the maximum appearance in the conformational ensemble of the **I**-state *i*) shifts in terms of temperature (**I**$_{max}^{model}$) as the TC-foldamer gets more compact but *ii*) does not change in terms of relative contribution [I]$_{max}^{model}$~70%. For shorter and thus least packed Trp-cage folds, contribution of **F** is smaller and that of **I** gets more important as temperature rises (*e.g.* E4). When the Trp-

cage core is more rigorously packed, E6-E11, the contribution of **F** is larger but decreases as temperature increases. While values of $[F]^{model}(T)$ decreases, those of $[I]^{model}(T)$ increases to some extent with rising temperature and thus **I**-sate is gradually replacing **F**-state: a subtle reorganization of the 3D-core structure of the Trp-cage fold is observed. Thus, in spite of the increasing temperature, the heat induced 3D-structure unfolding starts in the longer TC-foldamers only at a higher temperature. In other words, **I**-state is temporary but significantly helps to maintain the overall fold compactness of the TC-foldamer by delaying thermal unfolding of the protein. Therefore, the proportion of **U**-state, $[U]^{model}(T)$, remains low and the event of unfolding is postponed toward a higher temperature. In other words, the very same degree of T-induced unfolding (*e.g.* $[U]^{model}$~30%) is reached for longer TC-foldamers at a higher temperature. Expression all these in numbers is as follows: $[U]^{E4}$~30% at T=**30** °C, $[U]^{E5}$~30% at T=**45** °C, $[U]^{E11}$~30% at T=**50** °C.

Above, during the qualitative analysis, we have shown that the ^1L_b ECD band (*e.g.* Tyr (0→0) at 287 nm) changes as function of temperature, $[\Theta]^{\lambda,model}$~f(T) and present a T_{opt} value referring on the miniprotein fold compactness (**Table 2** and **Figure 5**). These values were proposed to use as fold compactness indicators and interpreted as the temperature where the overall 3D-fold compactness is the highest. Deconvolution now provides explicit conformational weights of the *T*-dependent **I**-state, $[I]^{model}(T)$. This measure also has a maximum as function of temperature resulting in I_{max}^{model} with a contribution of ~70% for most folded TC-miniproteins. Cross-correlation of T_{opt}^{model} values with those of $[I]_{max}^{model}$ - temperature (°C) where the percentage of **I**-state is maximum – is significant: R^2=0.945. Thus, these describe the same phenomenon and enforce the key role of **I**-state as the "alterego" of **F**-state and delaying temperature induced unfolding of these foldamers. This is a fascinating feature of TC-miniproteins in stabilizing their 3D-compactness even at elevated temperature, a possibility most probably present in other globular proteins as well. In conclusion, the quantitative analysis of the near-UV ECD spectral ensemble reveals that:

1) the intensity of the L_b electronic transition at 293 nm is associated with the compactness and relative orientation of a Trp↔Tyr sensor buried in the hydrophobic core.
2) **F-** and **I**-states in TC-foldamers have common structural features with subtle differences, though both give a C-type far-UV ECD spectrum.
3) Near-UV ECD spectral characteristics of the fully unfolded **U**-state is unambiguously given here, for the first time based on the joint analysis of hundred and more spectra.
4) T_{opt}^{model} values correlate significantly (R^2=0.945) with I_{max} temperature, the temperature where the relative contribution of **I**-state is maximum (~70%). This pinpoints that in fact **I**-state is the one which is responsible for the improved heat resistance of the *N*-terminal elongated miniproteins.
5) In principle, thermal unfolding occurs in two steps as **F**-state decreases but **I**-state initially increases. As **F** and **I** share a very similar overall scaffold with subtle differences[3] but the overall 3D-structure during an **F→I** transition is improving: TC-foldamer gets more compact though temperature is rising.

The present ensemble of ECD data reinforces not only the **F↔I↔U** three-state unfolding route but also enables generalization what was described for Tc5b.[4c] The molecular background of a non-linear or cooperative nature is provided and generalized here.

Conclusion

The **F**- to **U**-state non-linear unfolding mode of E5-E11 TC-proteins, as their α-helical segments are long enough, occurs *via* the key intermediate state, which efficiently retards thermal unfolding. Increasing internal motion as T rises reshapes the hydrophobic core **F** to **I** which becomes even tighter and thus less exposed to water up to 35-40 °C. The tighter the 3D-structure is, the more it prevents water penetration into the Trp-cage core as monitored by the near-UV ECD spectral changes.

Two alternative folding routes were proposed[11] where the predominant one is assuming an intermediate **I**-state also proposed by Zhou,[13] shown experimentally for the first time by Rovó *et al.*[4c] One folding path proposes the α-helix formed first while the second one suggests that key tertiary structure contacts first develop, followed by secondary structure evolution. Most probably, secondary structural elements (*e.g.* α-helix) are by themselves marginally stable only and therefore these have to be stabilized by key tertiary interaction. By analyzing poorly folded E0-E3 miniproteins having mainly unfolded even at *T*=5°C but Trp-cage folding is already detectable (X_{Fcage}~0.2-0.4), we give strong evidence that the second mechanism is the one which is the most probable. In line with our heteronuclear NMR results on Tc5b,[4c] here we generalize the folding concept of hydrophobic collapse: the second route proposed for TC-protein's folding.

Trp↔Tyr residue pair as sensor in proteins: a PDB analysis

The present comprehensive structure analysis of E0-E11 TC-miniproteins revealed how Trp↔Tyr residues are structurally oriented in a hydrophobic core and could be used as fold-sensors *via* their characteristic UV-ECD spectral shifts. The relative spatial orientation of these two sidechains is best characterized by *d* (Å), *Φ* (°), *Θ* (°) and *α* (°) descriptors[14] where *d* is the distance between the aromatic centers, *Φ* is the azimuth angle of Tyr aromatic center and *Θ is* the elevation angle from the plain of the phenol ring center above that of the indole plane of Trp and α is the angle between the two aromatic planes. Searching for Trp↔Tyr residue pairs in hundreds of proteins in the PDB database it was shown that their optimum separation is around *d*= 7Å, with a rapid decline at ~10Å. By applying a 9Å cutoff value, a total of 6499 interacting Trp↔Tyr pairs were collected from 3119 protein chains of not more than 25 % homology.[15] When filtering the NMR determined multiple copies out and retaining a single copy of them, a total of 1206 Trp↔Tyr pairs remained. In this high resolution structural dataset, only those interacting Trp↔Tyr pairs were selected which has no additional Tyr or Trp in their vicinity namely 0Tyr.Trp↔Tyr. (**STable 2**)

The 'descriptor distance' (ζ) is defined to be the minimum among the vectorial distances of the normalized descriptors calculated for each PDB structure with respect to E0-E11 miniprotein conformers:

$$\zeta = \sqrt{\left(d_{norm}^{PDB} - d_{norm}^{M}\right)^2 + \left(\Phi_{norm}^{PDB} - \Phi_{norm}^{M}\right)^2 + \left(\Theta_{norm}^{PDB} - \Theta_{norm}^{M}\right)^2 + \left(\alpha_{norm}^{PDB} - \alpha_{norm}^{M}\right)^2} \; .$$
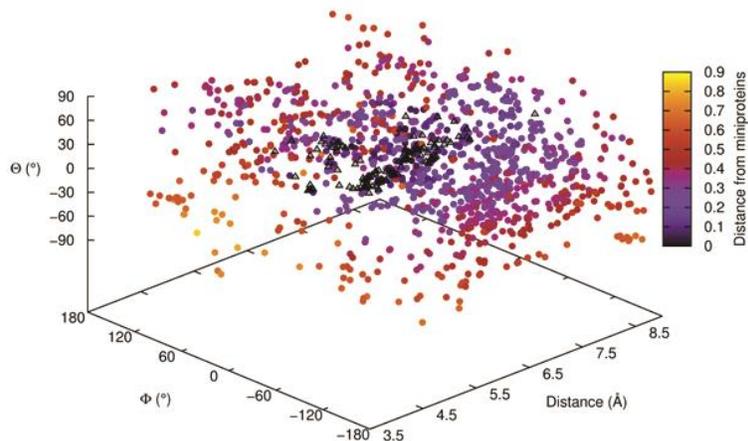
Normalization converts the different descriptors into the 0-1 range in line with the following restraints:

$$d_{norm}=(d-d_{min})/(d_{max}-d_{min}), \text{ where } d_{min}=3.5 \text{ Å}, d_{max}=9 \text{ Å}$$
$$\Phi_{norm}=\Phi/360°, \text{ since } 0°\leq \Theta \leq 360°$$
$$\Theta_{norm}=(\Theta+90°)/180°, \text{ since } -90° \leq \Theta \leq +90°$$
$$\alpha_{norm}=(\alpha/90°), \text{ since } 0° \leq \alpha \leq 90°$$

Figure 9. *Spatial distribution of the interacting Trp↔Tyr pairs of the 1206 proteins and 120 miniproteins plotted as function of d, Φ, Θ variables. Note that the ten best calculated NMR structures for each miniprotein model were plotted as black triangle (12 x 10 = 120 points) while those of proteins are colored according to their vectorial distance (ζ) from the closest miniprotein core-structure. (For a better visualization Φ was transformed from 0-360° to the equivalent -180-180° range.)*

The interacting Trp↔Tyr pairs of the TC-miniproteins are located in the central region of the *f(d, Φ, Θ)* space (**Figure 9**). If ζ is acceptably small ~0.2 (**SFigure 4**), then the Trp↔Tyr conformational properties of the miniproteins cover a significant part of the PDB conformational space. In total, the Trp↔Tyr conformational behavior of 223 proteins matches well with those of the miniproteins. Examples of selected proteins are also presented here to demonstrate that the orientations of the Trp↔Tyr pairs found in E0-E11 are similar to those of in proteins from PDB. In TC-miniproteins, Trp and Tyr are separated by 2 residues in the primary sequence also found in 53 proteins of PDB.
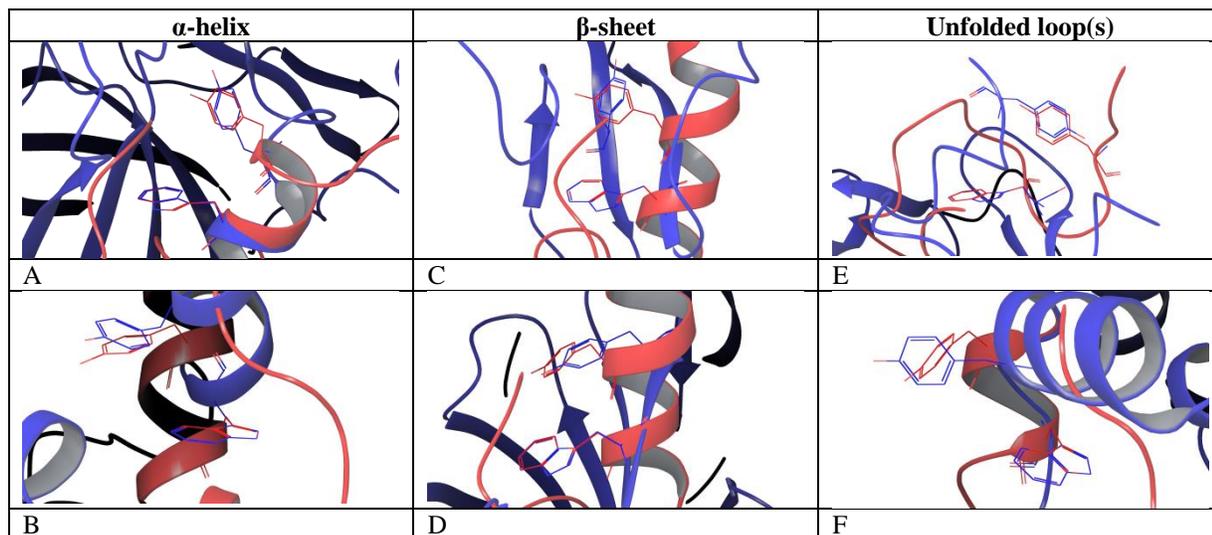
Figure 10. *Selected examples of Trp↔Tyr pairs in proteins (blue ribbon) mapped over the closest TC-miniprotein structure (red ribbon). Vectorial distances and the best superimposing TC-miniprotein type is given: A ζ=0.027 PDB:2UUY-A Trp(235)-Tyr(232) matching E2; B ζ=0.099 PDB:1I3Q-J Trp(18)-Tyr(21) matching E10; C ζ=0.084 PDB: 2DMV-C Trp(12)-Tyr(24) matching E11; D ζ=0.099 PDB:2F69-A Trp(120)-Tyr(122) matching E19; E ζ=0.049 PDB:1DQG-A Trp(41)-Tyr(134) matching E1; F ζ=0.085 PDB:1I3Q-A Trp(556)-Tyr(551) matching E2.*

The above used PBDselect comprises 3119 protein chains (homology < 25 %) in which among all proteinogenic amino acids, 354 777 in total, the abundance of Trp is 2.16 % and that of Tyr is 4.11 %. In this homology filtered database, an average protein chain length includes 113.8 residues: ~2.5±2 Trp and 4.7±4 Tyr residues are counted on. If these residues are located in the core part of the protein and are in close proximity to each other, the prerequisite of forming a near-UV sensitive chromophore, *i*) more than a single Trp↔Tyr pair per protein could be present and *ii*) additional one or two Tyr could surround each fold sensor Trp↔Tyr pairs. Therefore, the herein reported basic fold-sensing near-UV ECD spectra could become more complex. In spite of all these restrictions and forecasted complications, among these 3 119 proteins in total 1 206 Trp↔Tyr interacting aromatic residue pairs were assigned (see above) in a "moderate isolation" state (**STable 2**), cases where in principle the near-UV ECD spectra could perceive and report on both local and global unfolding of the protein. Several instances are spelled out below where Trp and Tyr sidechains are located in proximity ready for a 'direct', near-UV ECD based fold analysis.

Proteins or their domains listed below fulfill the above criteria and thus, their fold status could in principle be investigated by near-UV ECD spectroscopy. Selected examples listed are macromolecules involved in regulated cytoskeletal or spliceosomal assembly, in intracellular signaling, in protein degradation events, contribute to nucleic acid binding or to viral coat to name a few. The first one is the UBA domain of USP5 (ubiquitin *C*-terminal hydrolase 5; PDB: 2DAG), the protein that is required for normal proteosomal degradation, while the autoinhibitory GTPase binding domain (GBD) of WASP (PDB: 2K42) is a domain important for blocking WASP activation in controlling actin assembly. Subunit 2 of splicing factor 3B also termed as SAP 145 (PDB: 2DO5) is known for its function in spliceosomal assembly to edit pre-mRNAs contains an SAP domain at the *N*-terminus with close Trp and Tyr sidechains. The Pleckstrin Homology (PH) domain of AKT1/PKB (Protein Kinase B), a serine threonine kinase (PDB: 1UNQ), is involved in various cellular downstream events leading to *e.g.* cell survival and as such a target for cancer research. The DEP domain of Dsv-1 protein (PDB: 1FSH), a protein that mediates Wnt signal at the membrane and important in embryogenesis. The yet poorly characterized Zn-fingers (PDB: 2KGO and 2D9M) may bind nucleic acids and as such likely acts as a transcription factors. The Ebola virus glycoprotein (GP2 with its fusion loop at pH 5.5 but also at pH 7.0; PDB: 2LCY, 2LCZ, respectively) is important for virus-host fusion the hydrophobic core of which could also be analyzed by near-UV ECD spectroscopy. Furthermore, the pre-mRNA transcription enzyme from a eukaryote (RNA polymerase II 1I3Q), the SET7/9 a lysine methyltransferase (2F69), which typically methylates Lys of histone proteins and thus playing key roles in epigenetics and developmental processes is also to be mentioned. Itchy homolog E3 ubiquitin protein ligase (2DMV) regulates ubiquitination of several proteins and therefore leads to their degradation inhibitors (Tryptase-inhibitor 2UUY), a Kunitz/BPTI-type serine protease inhibitor from tick saliva receptors (mannose receptor 1DQG) also known as CD206, an important protein playing a role in receptor mediated endocytosis of glycoproteins by macrophages, all have Trp↔Tyr sensors enabling a 3d-fold study by ECD spectroscopy.

As illustrated by these representative proteins of biological significance, Trp↔Tyr pairs can be part of α-helix, β-sheet, unfolded loop(s), *etc.* of globular proteins (**Figure 10**), where the orientation of the Trp↔Tyr sensor is similar to that of the TC-miniprotein. As both sequential and spatial similarities of the herein described and 'calibrated' Trp↔Tyr sensor can be found in globular proteins, the near-UV ECD spectral characteristics provided here are of general interest, knowledge applicable for a wide range of proteins.

Methods

Experimental

Numbering of residues in this paper is according to E19 (Rovó et al 2014), [1]HGEGTFTSDL[11]SKQMEEEAVR[21]LYIQWLKEGG[31]PSSGRPPPS (**Scheme 1**). The number in the name indicates the length of the *N*-terminal elongation with respect to the reference 20-residue-long TC-miniprotein; E0: RLYIQWLKEGGPSSGRPPPS. E0 has no while E19 has 19 extra *N*-terminal residues, all similar to those of Ex4.

Peptide synthesis and purification

Peptides E0–E11 were prepared by solid-phase-peptide synthesis on MBHA resin using Boc technique. Resin was treated with liquid HF/dimethyl sulphide/p-cresol/p-thiocresol (86:6:4:2, vol/vol) at 0°C for 45 min. Once HF was removed, the free peptide was solubilized in 10% aqueous acetic acid, filtered and lyophilized subsequently. Crude peptides were purified by reverse-phase HPLC on a Phenomenex Jupiter C-18 column (21.2 × 250 mm) using a gradient of water/acetonitrile (solvent system was the following: 0.1% TFA in water and 0.1% TFA, 80% acetonitrile in water, gradient was: 0% → 40% B in 80 min, flow 3 ml/min and detection was made at 220 nm). Collected fractions were lyophilized, identities confirmed by a Finnigan TSQ 7000 tandem quadrupole mass spectroscopy equipped with an electrospray ion source.

Peptide expression and purification

Peptides E5, E10, E11 and E19 were prepared with recombinant biotechnology as published for miniproteins.[3,4c] cDNAs were ligated into a SacII and BamHI site of the pUBK2 vector and the plasmid encoding H10-Ub-Peptide was transformed into the *E. coli* strain BL21(DE3). Transformed cells were grown in LB medium containing 0.1 mg/ml Kanamycin and grown with shaking at 200 rpm 37 °C. Cells were induced at OD600 0.6 –0.8 by addition of IPTG to a final concentration of 1 mM. After 3 hour incubation at 37 °C cells were harvested and resuspended in lysis buffer (300 mM NaCl, 50 mM NaH_2PO_4 and 3 mM NaN_3) and lysed by sonication. Pellet was removed by centrifugation and the target protein was dissolved in the supernatant. All proteins were purified using 5 ml nickel-nitrilotriacetic acid (Ni-NTA) chromatography column: column equilibrated and washed with lysis buffer and then the fusion protein was eluted with the same lysis buffer which contains 250 mM imidazole. The eluated fractions were overnight dialysed into lysis buffer then the fusion proteins were digested with His-tagged yeast ubiquitin hydrolase (YUH) for 3 – 4 h and hydrolyzes monitored on SDS-PAGE. The ubiquitin and YUH were separated from the peptides by Ni-affinity chromatography. Peptides were further purified by reverse-phase HPLC on a C18 column using a gradient of water/acetonitrile (eluent A is 0.1% TFA in water and eluent B is 0.08% TFA, 80% acetonitrile in water). Collected fractions were lyophilized and their identities were confirmed by a Perkin–Elmer Sciex API2000 mass spectrometer equipped with an electrospray ion source.

### Electronic Circular Dichroism Spectroscopy

Far and near-UV ECD spectra were recorded on a Jasco J810 spectrophotometer using cuvettes with a path-length of 1.0 or 10 mm with protein concentration of 35-150 µmol. Typical spectral accumulation parameters were a scan rate of 50 nm/min with a 1 nm band-width and a 0.2 nm step resolution over the wave-length range 185-260 nm (farUV) and 250-320 nm (nearUV) with four scans averaged for each spectrum at temperatures ranging from 5 to 85 °C with 5 minute thermal equilibration for each 5 °C step. The temperature at the cell was controlled by a Peltier-type heating system. The spectrophotometer was equilibrated at each temperature for 4 minutes before acquisition. The solvent reference spectra were used as baselines which were automatically subtracted from the peptide spectra. The raw ellipticity data were

converted into mean residue molar ellipticity units ($[\Theta]_{MR}$, deg×cm^2×dmol^{-1}) for far-UV region and molar ellipticity ($[\Theta]$, deg×cm^2×dmol^{-1}) for near-UV region.

Descriptors calculation

Protein chains from the recent PDBSelect,[15] comprising a list of non-homologous proteins (25% homology, nsigma=3.0) was used to build the database of Trp/Tyr residue pairs.
The coordinate system based on the proposal by Singh & Thorton,[2] namely the orientation of CE2->CD2 atoms determine the *x*-axis, while the *y*-axis is perpendicular to it and goes through atom CD1 (**Figure 17**). With *x*- and *y*-, the *z*-axis forms a right handed *Cartesian* coordinate system. To describe the relative orientation of the residues –the vector from the origin to the center of the Tyr residue - the following 'polar'descriptors were used. The size of the vector considered as the distance of the residues. The polar angle Θ measures the angle between vector and the the *x,y*-plane while the polar angle Φ angle is calculated as the angle the vector to the Tyr center projected to the *x,y*-plane. *X*-axis is the 0° and angle counted counter-clockwise looking from the direction of the *z*-axis.
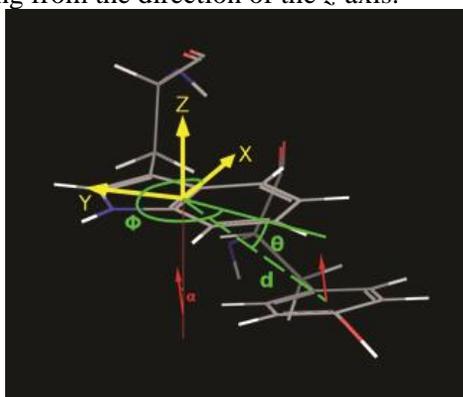
Figure 11. *Coordinate system and descriptors used for defining the relative orientation of the Trp↔Tyr residue pair as fold sensor. The X-axis resides on the line connecting the Trp CE2-CD2 atoms, the Y-axis is perpendicular to this line and goes through the Trp-CD1 atom. That defines the origin of the right-handed Cartesian coordinate system. The d distance of the Tyr is the distance of the origin and the centre point of the Tyr residue. The relative orientation is determined by the Φ azimuthal angle and Θ polar angle.*

References

[1] a) R. Kiss, Z. Bozóky, D. Kovács, G. Róna, P. Friedrich, P. Dvortsak, R. Weisemann, P. Tompa, A. Perczel *FEBS Lett.* **2008**, *582*, 2149-2154; b) R. Kiss, D. Kovács, P. Tompa, A. Perczel *Biochemistry* **2008** 47, 6936-6945; c) P. Tompa, *Structure and Function of Intrinsically Disordered Proteins*, Chapman and Hall/CRC, **2010**

[2] J. Singh, J.M. Thornton, *Atlas of Protein Side-Chain Interactions, Vols. I & II*, IRL press, Oxford, **1992**.

[3] P. Rovó, V. Farkas, P. Stráner, M. Szabó, Á. Jermendy, O. Hegyi, G.K. Tóth, A. Perczel, *Biochemistry* **2014**, *53*, 3540-3552.

[4] a) B. Barua, J.C. Lin, V. D. Williams, P. Kummler, J.W. Neidigh, N.H. Andersen, *Protein Eng., Des. Sel.* **2008**, *21*, 171-185; b) P. Rovó, V. Farkas, O. Hegyi, O. Szolomájer-Csikós, G.K.Tóth, A. Perczel, *J. Pept. Sci.* **2011**, *17*, 610−619; c) P. Rovó, P. Stráner, A. Láng, I. Bartha, K. Huszár, L. Nyitray, A. Perczel, *Chem. - Eur. J.* **2013**, *19*, 2628−2640.

[5] a) K. Kuwajima, E.P. Garvey, B.E. Finn, C.R. Matthews, S. Sugai, *Biochemistry* **1991**, *30*, 7693-7703; b) A. Perczel, M. Hollósi in *Circular Dichroism and the Conformational Analysis of Biomolecules*, (Ed. G.D. Fasman), Plenum Press, New York, **1996**, pp. 285-369; c) N.J. Greenfield, *Nat. Protoc.* **2006**, *1* 2891–2899; d) R.W.Woody in *Comprehensive Chiroptical Spectroscopy, Vol. 2* (Eds.: N. Berova, P L. Polavarapu, K. Nakanishi, R W. Woody), John Wiley & Sons, Inc., Hoboken, New Jersey, **2012**, pp. 475-497; e) R.W. Woody *Biomed. Spectrosc. Imaging* **2015**, *4*, 5-34.

[6] V. Farkas, B. Csordás, O. Hegyi, G.K. Tóth, A. Perczel, *Eur. J. Org. Chem.* **2013**, 3513-3522.

[7] a) W.J. Goux, T.M. Hooker, Jr., *J. Am. Chem. Soc.* **1980**, *102*, 7080-7087; b) N. Sreerama, M.C. Manning, M.E. Powers, J.-X. Zhang, D.P. Goldenberg, R.W. Woody, *Biochemistry* **1999**, *38*, 10814-10822; c) R. Li, Y. Nagai, M. Nagai *J. Inorg. Biochem.* **2000**, *82*, 93-101; d) Y. Jin, H. Sakurai, Y. Nagai, M. Nagai *Biopolymers* **2004**, *74*, 60-63; e) A. Vanhooren, A. Chedad, V. Farkas, Zs. Majer, M. Joniau, H. Van Dael, I. Hanssens *Proteins: Struct. Funct., Bioinf.* **2005**, *60*, 118-130; f) O.K. Gasymov, A.R. Abduragimov, B.J. Glasgow, *Anal. Biochem.* **2008**, *374*, 386–395; g) S. Nagatomo, M. Nagai, T. Ogura, T. Kitagawa *J. Phys. Chem. B* **2013**, *117*, 9343−9353; h) L.H. Fornander, B. Feng, T. Beke-Somfai, B. Nordén, *J. Phys. Chem. B* **2014**, *118*, 9247-9257; i) O.K. Gasymov, A.R. Abduragimov, B.J. Glasgow *J. Phys. Chem. B* **2014**, *118*, 986-995.

[8] E.H. Strickland, M. Wilchek, J. Horwitz, C. Billups, *J. Biol. Chem.* **1970**, *245*, 4168-4177.

[9] N.J. Greenfield *Nat. Protoc.* **2006**, *1*, 2527-2535.

[10] a) A. Perczel, M. Hollósi, G. Tusnády, G. Fasman, *Protein Eng.* **1991**, *4*, 669-679; b) A. Perczel, K. Park, G. Fasman, *Anal. Biochem.* **1992**, *203*, 83-93; c) I. Jákli, A. Perczel, *J. Pept. Sci.* **2009**, *15*, 738-752.

[11] a) J. Juraszek, P. Bolhuis, *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 15859-15864; b) J. Juraszek, P. Bolhuis, *Biophys. J.* **2008**, *95*, 4246-4257.

[12] J.W. Neidigh, R.M. Fesinmeyer, N.H. Andersen, *Nat. Struct. Biol.* 2002, *9*, 425-430.

[13] R. Zhou, *Proc. Natl. Acad. Sci. U. S. A.* 2003, *100*, 13280-13285.

[14] a) J. Singh, J.M. Thornton, *FEBS Lett.* **1985**, *191*, 1-6; b) L. Brocchieri, S. Karlin, *Proc. Natl. Acad. Sci. U. S. A.* **1994**, *91*, 9297-9301.

[15] S. Griep, U. Hobohm, *Nucleic Acids Res.* 2010, *38*, D318-D319