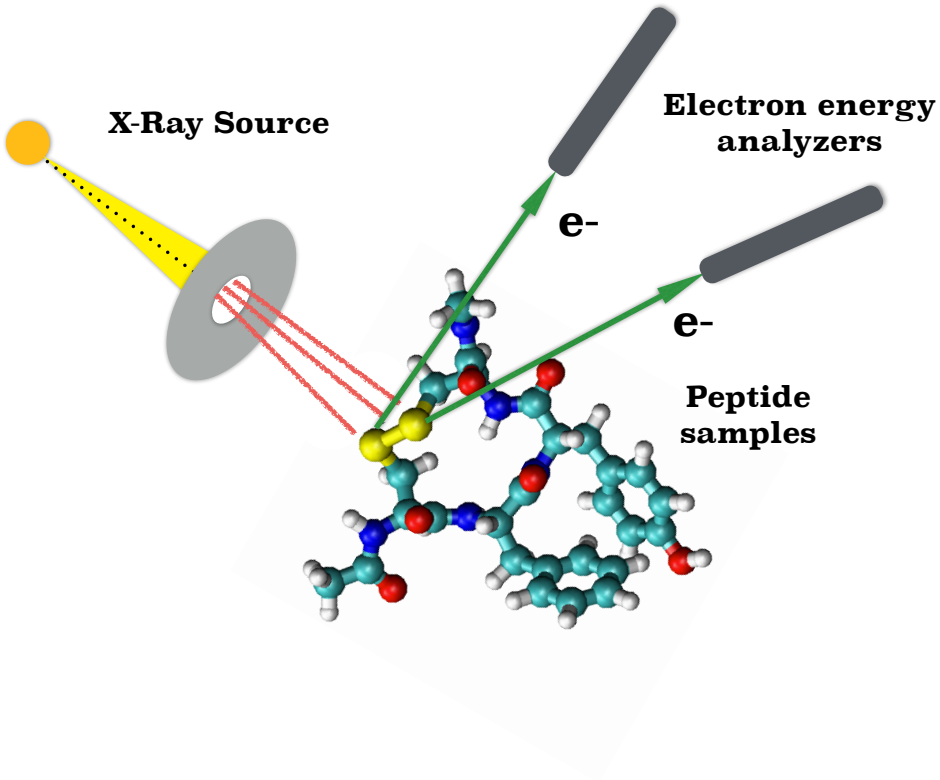
PCCP

Accepted Manuscript

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/pccp

Apparatus set-up for performing time-resolved synchrotron measurements of core level shifts in a model bio-molecule.



FULL ARTICLE www.rsc.org/xxxxxx | XXXXXXXX

Tetrapeptide unfolding dynamics followed with core-level spectroscopy: a first-principles approach

Simone Taioli*[a,b], Stefano Simonucci[c], Silvio a Beccara[a], and Marco Garavelli[e,f]

In this work we demonstrate that core level analysis is a powerful tool for disentangling the dynamics of a model polypeptide undergoing conformational changes in solution and disulphide bond formation. In particular, we present computer simulations within both initial and final state approximations of 1s sulphur core levels shifts (S1s CLS) of the CYFC tetrapeptide for different folding configurations. Using increasing levels of accuracy, from Hartree-Fock and density functional theory to configuration interaction via a multiscale algorithm able to reduce drastically the computational cost of electronic structure calculations, we find that distinct peptide arrangements present S1s CLS sizeably different (in excess of 0.5 eV) with respect to the reference disulfide bridge state. This approach, leading to experimentally detectable signals, may represent an alternative to other established spectroscopic techniques.

Introduction

The characterisation of dynamic properties of bio-molecules and proteins is fundamental for our understanding of their functions and for developing novel approaches to rational drug design and protein engineering.

In this regard, one of the most fruitful and widely used method is represented by two-dimensional (2D) NMR spectroscopy whose temporal resolution falls on the ms timescale, limited by the duration of the radio-frequency driving pulses.[1]

More recently, 2D spectroscopic techniques have been extended to the optical infrared (2DIR) and visible (2DVIS) regimes.[2] In particular, 2DIR spectroscopy uses ultra-short laser pulses to investigate structure and sub-picosecond dynamical processes in biological systems, such as chemical exchange, vibrational population transfer, and molecular reorientation of DNA oligomers and backbone amid groups[3]. While this approach represents a true step forward for real-time dynamics characterisation with respect to standard linear absorption or NMR spectroscopies, 2DIR of complex biomolecules may yield highly congested spectra with overlapping signals that require intensive isotope labelling to obtain unambiguous identification. 2DVIS spectroscopy achieves an even higher temporal (fs timescale) and spectral resolution, relying on fs laser pulses to track electronic transitions of visible absorbing multi-chromophoric systems, such as light harvesting complexes and photosynthetic reaction centers.[4-5]

Further extension of 2D spectroscopy to ultraviolet spectral regime[6] (2DUV) would enable to follow the molecular dynamics of complex UV absorbing biosystems, notably DNA, RNA, and proteins[7]. Recent computational studies have confirmed that 2DUV spectroscopy has a great potential both in the structural determination of these biomolecules[8-9] and in tracking conformational dynamics of model polypeptides[10]. However, only preliminary 2D experimental studies, with limited pulse bandwidths, have been reported in the UV so far[6] due to technical difficulties in achieving interferometric stability, ultra-broadband pulses and a good signal-to-noise ratio at these energies. Furthermore, to obtain signals showing sensitivity to conformational dynamics one should perform two-color pump-probe experiments recording contributions from different chromophores, whose interpretation can be exceedingly complex. Thus, the capacity of this technique to fully resolve folding/unfolding processes remains disputed.

To overcome these issues, in this work we propose a different approach for disentangling conformational changes in a model polypeptide that undergoes conformational changes in solution, aiming at demonstrating that core electron binding energies (BE) measurements may complement the spectroscopic methods discussed above. In particular, we demonstrate that signals recorded from sulphur atomic sites in the case of disulphide-bond formation of a folding CYFC (cysteine−phenylalanine−tyrosine−cysteine) are a fingerprint of the distinct geometric configurations visited during conformational changes and can thus be used to track dynamics both in gas-phase and solution.

Time-resolved core-level photoelectron spectroscopy can be in principle advantageous or, at least, competitive with respect to NMR and other techniques used to study conformational dynamics of proteins, such as Förster resonance energy transfer (FRET) and pulsed electron paramagnetic resonance (EPR or DEER). Indeed, NMR requires long acquisition times along with using expensive set-up, FRET needs fluorescent chromophores to be able to detect molecular interactions and therefore to provide information about protein conformation, and finally DEER can be applied only to paramagnetic compounds.

Core-level photoelectron spectroscopy is a rather general approach to probe the electronic structure, applicable to a wide variety of physical systems, and providing a chemical selective method by tuning the incoming x-ray beam energy. Furthermore, is a rather sensitive technique, able to detect a desired signal above the noise level from atoms or impurities present at concentration as low as 5%. Of course, the number of electrons emitted from a given species is related to its concentration within the system and the majority of electrons detected come from within few inelastic mean free paths. Indeed, photoelectron spectroscopy based approaches are known to be surface sensitive, while common bulk spectroscopic techniques such as NMR cannot be effectively used for surface studies as they cannot

provide good sensitivity levels. Finally, effect of light polarisation leading to magnetic dichroism can be studied by core-level photoelectron spectroscopy.
Time resolution of this technique is comparable to that of pump-probe experiments as isolated sub-femtosecond soft-X-ray pulses can be generated, core hole is created within less than a femtosecond and photo-electron wave packets sampling can be attained within attosecond, while its subsequent decay can be observed on a few-femtosecond timescale. Thus, this technique offers attosecond resolution[11].
Finally, core-electron spectroscopy could be performed on any of the atoms and at each configuration both in our model and actual experiments, as energy tunability is one of the most important advantages that can be achieved at synchrotron light sources and used by this approach. Besides this, high intensity of the photon beam can be achieved in such facilities obtaining an overall better sensitivity.

To demonstrate the potentiality of core electron spectroscopy to investigate folding dynamics, we apply Hartree-Fock (HF), Density Functional Theory (DFT) and Configuration Interaction (CI) to track disulphide bond breaking/forming in a model CYFC peptide by calculating BE and excitonic spectra of sulphur atoms participating into the bridge forming between cysteine residues in solution. This process plays a major role in the stabilization of protein folding mechanisms. We show that BE can be qualitatively obtained by mean-field methods, such as HF and DFT, but one needs to use approaches beyond mean-field, notably CI, to reach chemical accuracy.

Results and discussion

To accomplish this task, we selected 12 configurations from the most probable folding trajectory for the CFYC, generated by means of the Dominant Reaction Pathways (DRP) approach[12], in explicit solvent. DRP yields the most probable unfolding pathway for macromolecules between given initial and final conformations, and it is based on a variational solution of the Fokker-Planck equation in the path integral representation. Details of the approach are given in the Computational methods section. The peptide dynamics is labelled by the S-S distance along the optimal trajectory (see Fig. 1). After disulphide bond cleaving (which can be due for example to interaction with electromagnetic fields), in this work we assume that both sulfur radicals are immediately saturated by hydrogen atoms from close-by solvent molecules and the saturation is fast compared to the unfolding dynamics. Thus, immediately after disulphide bridge breaking we introduce hydrogen atoms to saturate sulfur atoms. Note that, before complete separation, the S-S distance increases and oscillates between 7 and 9 Å (see Fig. 2, red curve) due to solvent stochastic forces. This is responsible for the oscillating behaviour of the CLS in Fig. 2a, which is a further demonstration of the technique sensitivity to the interaction between solvent and thermal bath with the peptide.
In this regard, the initial and final geometries are, respectively, the T-stacked configuration characterised by disulfide bridge (see Fig. 1a) and the unfolded state where aromatic rings are far each other (see Fig. 1n) and sulphur-sulphur (S-S) bond is broken. Initial and final configurations have been selected as the global minima of 50 ns classical molecular dynamics simulations of the folded (S-S bonded) and unfolded CYFC. The peptide dynamics is labelled by the S-S distance along the optimal trajectory (see Fig. 1). Before a complete breaking, the S-S separation increases and oscillates between 7 and 9 Å (see Fig. 2, red curve) due to solvent stochastic forces.
Our goal is to determine if core level (CL) analysis is able to disentangle between this oscillating behavior. In this regard, photo-excitation from atomic CL in a molecule consists in the creation of an electron-hole pair by an incident photon. Distinct information is derived by considering the energy range whereby the electron is imparted sufficient energy either to be emitted from the molecule to the continuum (photoemission or XPS) or to be excited to empty orbitals in metastable states (absorption or XAS)[13].

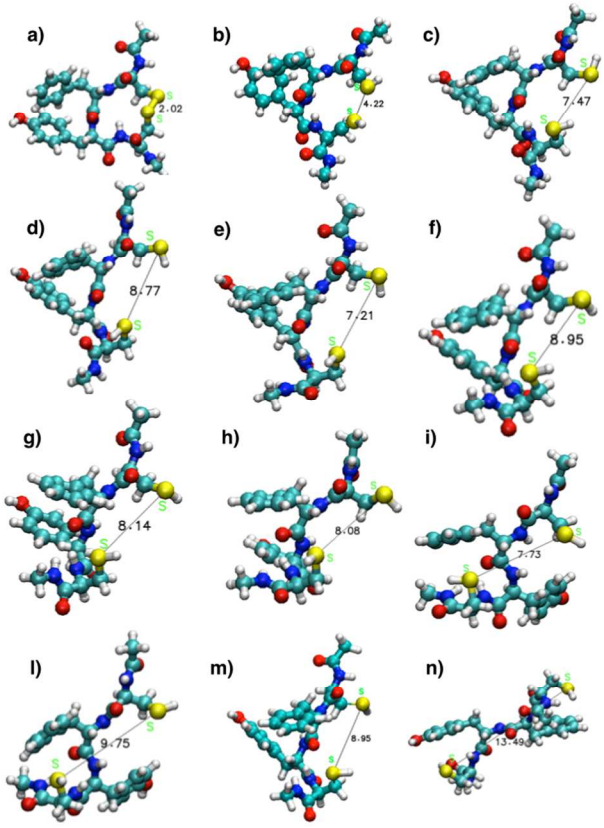


Figure. 1 Snapshots of the variational optimal CYFC folding path used in the BE calculations, showing the formed (a) and broken (n) disulfide bridge. S-S distance (Å) is the representative coordinate.

The most fruitful approach to electron spectroscopy has been to understand measured spectra in terms of differences, called core level shifts (CLS), between the BE ε_b of a given state and some reference BE ε_b^{ref} that can be chosen arbitrarily (see the Computational methods section for details):

$$E_{CLS} = \varepsilon_b - \varepsilon_b^{ref} \quad (1)$$

Generally, measured CLS are in qualitative agreement with HF and DFT mean-field calculations within a one-electron picture. Nevertheless, they result from a combination of many-body correlation effects observed through the appearance of energy shifts taking place in the initial (IS) and final states (FS) of the photoemission process. CLS were obtained by selecting the initial disulphide bonding state BE as a reference value, using both IS

and FS approximations.

IS effects are associated with the mean ground state electrostatic potential: depending on the chemical environment, and hence on the charge on the atom, ε_b can increase or decrease as compared to the free atom. Within this approximation:

i) the *N-1* electron system, left behind by photoemission, remains frozen in its original ground state;

ii) the core-hole remains localized at the atomic site where was initially created.

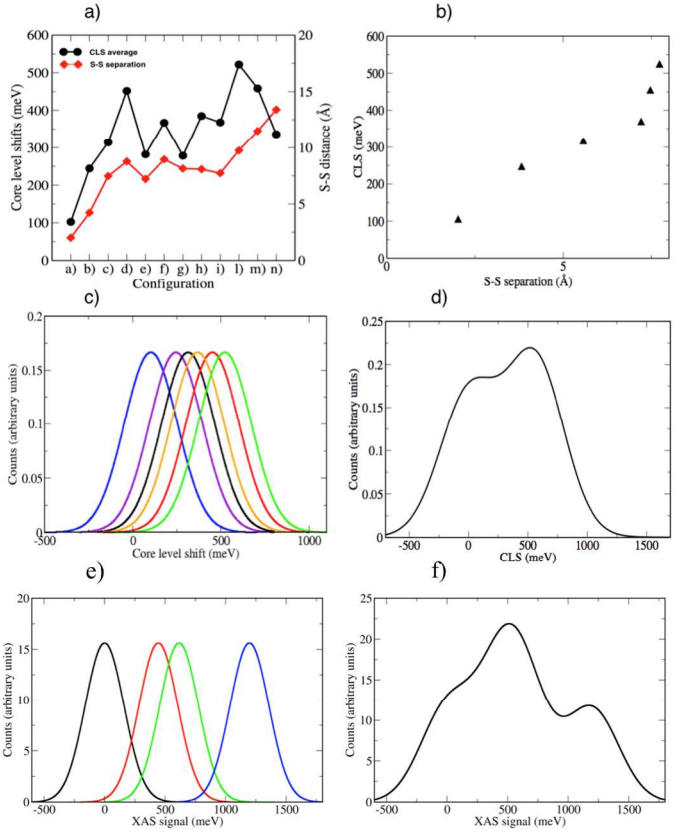


Figure. 2: a) S-S distance (red, Å) and averaged (see text for details) CLS (black, meV) for the 12 configurations as by Fig. 1. b) CLS (meV) vs. S-S separation (Å) for 6 CFYC configurations. c) CLS signal progression (0=reference CLS) for 6 CFYC configurations, from the initial disulphide bridge configuration (blue curve) to the final geometry with S-S broken bond (green curve) assuming the same population for all configurations during transient. d) Theoretical CLS Voight profile spectrum (0=reference CLS) obtained by convolution of a Gaussian (FWHM=0.6 eV) and a Lorentzian (FWHM=0.2 eV) function assuming that the initial and final states of the trajectory are the most populated at equilibrium. e) XAS signals (0=reference XAS) for the initial and final configurations. f) XAS Voight profile spectrum (0=reference XAS) obtained by using he same parameters as in d).

FS effects are associated with the response of the electronic cloud of the system to the core-hole potential. This relaxation causes a decrease in the total energy of the system which is taken off by the photoelectron, thereby reducing ε_b. Using definition 1, cancellation of errors due to inaccurate treatment of the electronic correlation at these stages may provide a quantitative interpretation of the measured electron spectra.

In mean-field theories, such as HF or DFT, IS approximation consists in identifying core electron BE as the HF and Kohn-Sham eigenvalues obtained via a self-consistent charge density calculation.

An all-electron approach was employed to perform HF calculations assuming that the photon energy is much higher than the photoionisation threshold, so that the photoelectron leaves the core-hole region in a short time compared to the surrounding electronic cloud response. HF (and CI) calculations have been carried out using our own code suite SURPRISES,[14-15] optimized for performing CL analysis. We used a 6-31G* basis set, containing (*s,p,d*)-type variationally optimized Hermite gaussian functions. Atomic charge density was chosen for the initial step in all HF calculations. From the HF solution, one obtains a complete set of orbitals among which the lowest in energy are chosen to construct the one-determinant wave function that represents the ground state of the system.

DFT calculations were carried out within IS approximation by relaxing the valence charge density as response to the core-hole perturbation (see Computational methods section for details). CYFC configurations for DFT calculations were kept fixed to those obtained by DRP. CLS obtained by HF and DFT within IS approximation are reported in Tab.1. With respect to the folded reference state, CLS present sizeable dependence on configuration, with values in excess of 1 eV between native and unfolded states.

Table 1 S1s CLS (meV) within the IS approximation from HF and DFT (GW) calculations for different CYFC arrangements as by Fig. 1.

FC	S1 (HF)	S2 (HF)	S1 (GW)	S2 (GW)
a)	167	Ref.	130	Ref.
b)	562	1183	583	1110
c)	530	714	532	710
d)	864	1124	813	1016
e)	785	1092	768	1087
f)	420	1264	488	1219
g)	340	942	287	1038
h)	733	924	680	1043
i)	810	1189	910	1119
l)	450	1146	416	1098
m)	573	913	600	850
n)	778	1137	759	1123

The second approach to CLS analysis is represented by FS approximation. Within this framework relaxation of the remaining *N-1* electrons to screen the Coulomb-hole charge (including the exchange) at both intra-atomic and extra-atomic level leads the system to a new energy state, lower than the IS one by the relaxation energy gain. To assess the impact of relaxation on CLS, we performed HF calculations of core excited states within FS approximation. We found out that, even including only static correlation, FS approximation plays a major role in lowering CLS, as one can see in Tab. 2.

Finally, to improve the quality of the one-determinant wave function representation and obtain accurate CLS consistent with

experimental results, we used correlated methods beyond mean-field, notably CI, with all the electrons free to relax after the core electron removal. Dynamical electron screening is responsible for a whole range of effects, modifying the energy of both the ground *N* and photo-excited *N-1* electron systems, with an overall effect on the BE. However, while using all the orbitals of a complete set leads in principle to an exact expansion, the rate of convergence is very slow as:

i) many Slater determinants are necessary to reach chemical accuracy,

ii) the scaling with system size ($\sim M^6$) is particularly unfavourable for proteins, where the number of atoms *M* can be large. Recently, Taioli *et al.* [14-15] proposed a theoretical method for overcoming the scaling issue in CI calculations of electron spectra from molecules and solids.

The first step of this approach, whose details are reported in the computational methods section, consists in identifying within the system a cluster of atoms large enough to reproduce accurately the properties of interest. In our calculations, the cluster size (20 atoms in total) was increased until S1s CLS ceased to vary and, eventually, included the 10 atoms closely surrounding each S atom in every configuration visited by CYFC during S-S bond breaking/forming. The cluster choice reflects the assumption that CLS are affected only by the local density of states (DOS) as the range of interactions spans few atomic diameters. The remaining part of the system has the effect of an electron bath, shifting up or down CL BE.

Electronic structure calculations were approached by a multiscale algorithm, using CI with single-double excitations from the HF single determinant ground state for the cluster and a computationally affordable mean-field approach for the embedding bath. The CI active space is represented by the reduced manifold spanned by the cluster's orbitals obtained by the orthogonalisation procedure explained in the Computational methods section.

In DFT, electrons are removed from the core and placed into the valence. A corresponding core excited ionic pseudo-potential was generated "on the fly" during the self-consistent calculations.[16] In Tab. 2 we report CLS from HF, DFT (for both GW and LDA exchange correlation potentials) and CI within the FS approximation. The general trend is that HF overestimates CLS, while DFT and CI are in good agreement. CLS of this size can be clearly resolved by electron spectrometers, as S1s FWHM is around 0.6 eV and electron spectra with a total energy resolution (monochromator plus electron spectrometer) considerably smaller than the linewidth of the core/valence level investigated are nowadays obtainable. It turns out that averaging the 1s sulfur CLS of the two sulfur atoms at each of the 12 configurations selected along the trajectory, using the same level of theory, be it HF, CI or DFT, one obtains a quantity that can be directly related to the sulfur-sulfur distance via a one-to-one mapping.

This is particularly evident at CI level. In Fig. 2a) we plot the average CLS at CI level of theory along with the S-S separation as a function of the geometrical configuration. If our guess on the relation between these two observables was true, by plotting one quantity against the other one should clearly see this trend. Thus, in Fig. 2b) we report CLS for 6 configurations, those having the most appreciable difference in the sulfur-sulfur bond distance among the 12 investigated in this work. While Fig. 2b) can be simply obtained by Fig. 2a) and does not provide more information, by plotting CLS for those configurations having the most appreciable difference in the sulfur-sulfur bond distance as a function of bond length, one can clearly see this mapping and further appreciate the method sensitivity.

Table 2 S1s CLS (meV) with respect to the reference bridged state (Ref.) within the FS approximation from HF and GW-DFT (LDA results in parenthesis) for different CYFC arrangements as by Fig. 1. Furthermore, we report the CLS at CI level of theory for the initial (a-water) and final (n-water) configurations including water within a radius of 4 and 5 Å from the sulfur atoms with respect to the reference bridged state in the solvent.

FC	S1 (HF)	S2 (HF)	S1 (CI)	S2 (CI)	S1 (GW)	S2 (GW)	S1 (XAS)	S2 (XAS)
a)	194	Ref.	205	Ref.	183 (190)	Ref.	444	Ref.
b)	163	384	121	368	100 (90)	340 (330)		
c)	280	450	260	368	242	342		
d)	319	474	380	524	369	500		
e)	263	370	188	376	151	362		
f)	192	647	182	552	175	520		
g)	326	370	252	305	241	298		
h)	486	517	376	394	366	381		
i)	323	526	280	456	231	407		
l)	584	718	485	559	419	516		
m)	488	627	363	557	323	475		
n)	286	518	271	401	239	393	615	1196
a-water 4 Å			258	Ref.				
n-water 4 Å			342	513				
n-water 5 Å			392	542				

In Fig 2c) we report CLS theoretical spectra progression for different CYFC configurations, from the initial reference configuration forming a disulphide bridge to the final S-S broken bond geometry assuming that configurations are equally populated during the transient. In Fig. 2d) we report a theoretical CLS Voight profile, taking into account the experimental broadening and assuming that the initial and final states of the trajectory are mostly populated (steady conditions). Finally, in Fig. 2e) and f) we plot the XAS signals for the initial and final configurations of the peptide, using the same resolution as for CLS (see Fig. 2c)-d).

We clarify that our unfolded final reference state has been obtained as a local minimum of a long 20 ns molecular dynamics simulation after reducing the peptide with hydrogen atoms and, thus, is representative of a rather stable configuration. We note that many final stable representations, which possibly lower the

resolution of the CLS peak in the final state, could be present.

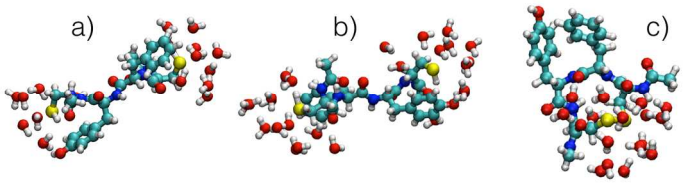


Figure. 3 Initial (a,b) and final (c) variationally optimized configurations of CYFC with explicit solvent within 4 (a) and 5 (b) Å radial distance from S atoms.

However, in this particular case, the sulfur-sulfur distance would not change significantly after reaching the final state, and CLS are totally uncorrelated at this point as clear from Fig. 2. Thus, in principle one should analyse the oscillating S-S distance in the final configurations but due to the one-to-one mapping we do not expect a significant broadening of the CLS peaks.

Additionally, we tested the effect of solvent on CLS by performing computer simulations including water within 4 and 5 Å radial distance from S atoms, as shown in Fig. 3. The effect of solvent at all level of theory is to shift almost rigidly BE increasing CLS of about 0.1 eV with respect to gas-phase, thus not affecting the relative trend of the signal. In Tab. 2 we report sulfur CLS obtained by adding water molecules within a distance of 4 and 5 Å from sulphur atoms at CI level of theory.

Furthermore, in order to check how CLS are affected by quantum mechanical (QM) optimization of the selected configurations obtained by classical molecular mechanics (MM) in connection with DRP, we performed QM relaxation of the initial and final configurations at CISD level of theory. Quantum mechanical optimized structures can be found in the DFT calculation section (Fig. 4). We find that CLS from these QM optimized geometries are practically indistinguishable, within the accuracy of our CI method, from those obtained in the MM configurations, differing only of some units of meV.

Finally, we performed the calculation of CFYC absorption spectrum (XAS) by probing the unoccupied DOS for different configurations. Close to the photoionisation threshold, the kinetic energy of the emitted photoelectron is so low to remain in the core-hole region for long time as compared to the surrounding electron response. Escaping electron and remaining hole binds into exciton, lowering the energy of the system. We characterised the excitation spectrum by removing the S1s core electron to the lowest energy singlet excited state, using the previous multiscale CI procedure for the initial and final configurations. The results, reported in Tab.2, show that even XAS signals are strongly dependent on the conformational changes of the peptide and thus even this spectroscopy could be used in principle to track conformational changes. Exciton BE are in the range of 1.5 to 3 eV.

Conclusions

In conclusion, we showed that CL analysis has the potential to characterize S-S bond breaking/forming in a model peptide. While we discussed this specific test-case, we believe that our method is totally general and can be used to disentangle folding dynamics in more complex biological systems. Thus, this work is aimed at stimulating experimental work by time-resolved high-resolution synchrotron radiation measurements of CLS in folding biological assemblies to find correlations between structural and functional properties in proteins.

Computational methods

Optimal folding path calculation

The most probable finite temperature pathway for the CFYC unfolding has been obtained by means of the Dominant Reaction Pathways (DRP) approach in explicit solvent. DRP is a variational boundary value method based on the minimisation of a functional of the path. Both the initial and the final conformations need to be given as an input. The derivation of the functional suitable for explicit solvent simulations is given in ref. [12]. The functional used in our work is:

$$\int_0^t d\tau \sum_{i=1}^{N} \frac{1}{\gamma_i m_i} \left| \boldsymbol{F}_i^{bias}(\bar{\boldsymbol{X}}) \right|^2$$

Here $|F_i^{bias}|^2$ is the bias force modulus, γ_i is the friction coefficient and m_i is the mass of the i-th atom. The integral is carried out up to time t. The bias force is only applied to peptide atoms. A swarm of 48 trajectories, connecting the initial and final conformations, was generated by means of ratchet and pawl molecular dynamics (rMD), and the functional was calculated for each trajectory. The pathway minimizing the functional is the most probable. The initial and final configurations were taken from ref. [10] and were chosen as a representative simple model of the initial and final state of a protein folding process. The sulphur-sulphur bond in the closed configuration enables the process to be started by a laser flash in an experimental setup. The sulphur atoms appear in their reduced form in the simulations due to the fact that molecular topologies cannot be changed in the course of the simulations employing empirical force fields. The initial and final conformations were locally minimised with the AMBER99SB-ILDN force field in explicit TIP3P water before starting the Dominant Reaction Pathways (DRP) simulations, attaining a locally stable configuration, and then equilibrated for 100 ps, with the same force field

Along this optimal pathway, 12 snapshots were selected to represent the CYFC unfolding process as reported in Fig. 1. On this specific set of configurations, characterized by different stacking of the aromatic side chain, we performed first-principles calculations of the 1s core-level electron binding energy for the two sulphur atoms (S1 and S2) participating in the bridge.

Core level shift calculations

Definition of binding energy

Photo-excitation from atomic core levels within a molecule consists in the creation, by an incident photon of energy *hv*, of an electron-hole pair. Generally, measured electron spectra are in qualitative agreement with HF and DFT mean-field calculations. However, electron correlation effects are easily observed through the appearance of energy shifts due both to final state effects and changes in the surrounding chemical environment with respect to expectations from the one-electron picture.[12] Thus, to obtain accurate theoretical spectra one needs to go beyond mean-field

and use perturbative or CI methods.

Denoting by E_0^N the ground state energy of a *N*-particle system, and by E_0^{N-1} the energy of the system after removing a particle from a single particle state *n*, conservation of energy for the photoemission process is given, within a one-electron picture, by:

$$h\nu + E_0^N = \varepsilon_{kin} + E_n^{N-1} \quad (3)$$

where ε_{kin} is the kinetic energy of the emitted electron. The photoelectron binding energy is defined as:

$$\varepsilon_b = h\nu - \varepsilon_{kin} = E_n^{N-1} - E_0^N \quad (4)$$

where all energies are referred in our case to the vacuum level.

However, dynamical electron screening cannot be indeed neglected, being responsible for a whole range of effects, such as: i) relaxation of the remaining *N-1* electrons to screen the hole charge (core-hole relaxation); this process involves both intra-atomic (inside the atom where the photoelectron has been created) and extra-atomic charge and leads the system to a new energy state, lower than E_n^{N-1} by an amount E_{relax};

ii) correlation among electrons: it affects both initial and final states. Correlation is described via the so-called *Coulomb hole* (reduced charge density, in the electron neighbourhood, due to Coulomb interaction) and the so-called *exchange hole* or *Fermi hole* (reduced density of electrons with the same spin, in the electron neighbourhood, because of the Pauli principle). Correlation modifies the energy of both the *N* and *N-1* electron systems, with an overall effect on the binding energy described by E_{corr}. As a consequence, the photoelectron binding energy in the final state approximation becomes:

$$\varepsilon_b = h\nu - \varepsilon_{kin} - \delta E_{relax} + \delta E_{corr} \quad (5)$$

In order to obtain core level spectra consistent with experimental results, we used correlated methods beyond the initial state approximation, notably CI, with all the electrons free to relax after the core electron removal.

The program SURPRISES

CLS calculations were performed using the in-house code suite SURPRISES.[14-15] This code provides an extension of Fano's resonant multichannel scattering,[17] allowing the construction of the continuum wavefunction of the emitted electron with appropriate boundary conditions, including the main correlation effects.

The first step of this general theoretical framework for interpreting electron spectra in molecules and solids implements a multiscale approach consisting in identifying a cluster of atoms, embedded in the system under investigation, large enough to reproduce accurately the property of interest. In our case, the relevant observables are represented by the 1s core levels of the two sulphur atoms (S1s) participating into the bond formation/disruption. The remaining part of the system has the effect of an electron bath shifting up or down core level binding energies. Cluster and bath will be treated at different level of accuracy, using a post-HF method, such as CI, for the former while a computationally more affordable mean-field approach for the latter. Indeed, the bottleneck of many ab-initio approaches to electronic structure in biophysics applications is the size of the functional space used to represent the wavefunctions, particularly in CI calculations. Our idea stems from considering that electron correlations to be included in core level analysis have a range of a few atomic diameters. This short interaction range allows one to use a reasonably small number of atoms to calculate CLS, which are affected only by the local density of states. In our calculation, the cluster size was increased until S1s CLS ceased to vary and, finally, included 10 atoms closely surrounding each sulfur atom (totally 20 atoms) in every configuration visited by the tetrapeptide during folding.

Special care has to be paid to quantify the interactions (bonds) of this cluster with the covalently bound surrounding environment. Such interactions cannot be simply neglected and these are the source of the unfavourable scaling with system size. In order to overcome this issue, we need to find a rigorous partitioning of the two functional subspaces generated by the Hermite gaussian functions (HGF) centered on the nuclei of the cluster and of the bath, respectively. For example, in the specific case of CFYC we have a complete manifold spanned by 498 HGF, of which, originally, 196 are centered on the cluster's nuclei. To orthogonalize cluster and bath manifolds, we performed initially an HF self-consistent calculation of the system (including cluster and embedding environment), obtaining a number of eigenvectors and eigenvalues (N_{tot}) corresponding to bi-occupied (N_b), mono-occupied or ligand (N_m) and virtual (N_v) orbitals of the system, such that $N_{tot} = N_b + N_m + N_v$.

HF orbitals, representing sulfur core levels, are of course delocalized on the entire tetrapeptide, thus the functional space spanned by cluster's HGF is not orthogonal to the environment manifold. We define two projectors for the bi-occupied and (ligand+virtual) orbitals of the cluster as follows:

$$P_b = \sum_{ij}^{N_b^{cluster}} |g_i\rangle S_{ij}^{-1} \langle g_j| P_v = \sum_{ij}^{N_v^{cluster}} |g_i\rangle S_{ij}^{-1} \langle g_j| \quad (6)$$

where $N_b^{cluster}$, $N_v^{cluster}$, and $N_m^{cluster}$ are the bi-occupied, ligand and virtual orbitals of the cluster, such that:

$$N_{tot}^{cluster} = N_b^{cluster} + N_m^{cluster} + N_v^{cluster} \quad (7)$$

and $S_{ij} = \langle g_i | g_j \rangle$ is the overlap matrix between HGF.

After projection of the entire system HF orbitals onto the cluster's 196-dimensional manifold by means of the projectors of Eq. 6, we separately diagonalize these operators. Finally, eigenvectors corresponding to eigenvalues equal to zero (one) can be safely attributed to orbitals external to (internal to) the cluster, while intermediate values (between 0 and 1) correspond to bonding orbitals to be included in the CI procedure. Bi-occupied and external orbitals were clamped down and did not have any role in the further CI procedure. Slater determinants for CI calculations were built using only orbitals having a not negligible (eigenvalues > 10^{-3}) projection on the cluster's functional space. In our case such a manifold has dimensions ranging between 234 and 239 depending on the configuration visited by the peptide (we remember that initial cluster's manifold is 196-dimensional, while the total functional space is 498-dimensional). Without loss of accuracy, we are thus able to lower the computational scaling by performing single and double excitations within the full functional space (full SDCI) spanned by the orthogonalized

cluster's orbitals, taking appropriate linear combinations of such Slater determinants (usually above a few million for this model tetrapeptide) to form spin-adapted singlet configurations.

DFT calculations

CLS were calculated within DFT using the total-energy and molecular dynamics VASP code,[18-20] with an efficient extrapolation of the electronic charge density.[21] Ion-electron interaction has been described using the Projector Augmented Wave (PAW) technique[22] with single-particle orbitals expanded in plane waves with a cut-off of 400 eV. GW exchange-correlation potential[23-24] with inclusion of dispersion interactions has been used to treat electron-electron Coulomb interaction, as GW PAW+VdW might be more accurate for calculating excited state properties. However, tests were performed by adopting LDA exchange-correlation potential finding no appreciable changes in CLS. In all DFT simulations, electronic and ionic temperature was fixed at 300 K with a corresponding Fermi smearing of the electronic density. Brillouin zone was sampled at the Γ-point only. CFYC was included in a supercell of 25X25X25 $Å^3$ to avoid spurious interactions among periodic images. Finally, DFT was used to optimize the sulfur-sulfur bond length in the open and closed configurations. In this case, only the position of sulfur atoms was left free to change along with close-by atoms up to the third star of neighbours, while the rest of the atoms was kept fixed. In the closed configuration sulfur-sulfur bond increased of 0.07 Å with respect to classical value, while 0.08 Å in the final configuration (see Fig. 4).

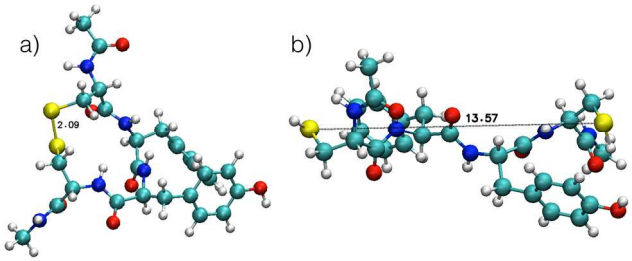


Figure. 4 Snapshots of the DFT optimized CYFC initial (a) and final (b) configurations. S-S distance (Å) is the representative coordinate.

Acknowledgments

The authors acknowledge useful discussions with dr. I. Rivalta (CNRS-Lyon) and prof. S. Mukamel (UC-Irvine). The research leading to these results has received funding from the European Union Seventh Framework Programme under grant agreement n. 604391 Graphene Flagship. S.T. acknowledges support by Istituto Nazionale di Fisica Nucleare through the "Supercalcolo" agreement with Bruno Kessler Foundation S.T. acknowledges economical support from the European Science Foundation under the INTELBIOMAT Exchange Grant "Interdisciplinary Approaches to Functional Electronic and Biological Materials" and the Bruno Kessler Foundation for providing economical support through the "research mobility scheme" under which this work has been accomplished. Furthermore, S.T. gratefully acknowledges the Institute of Advanced Studies in Bologna for the support given under his ISA research fellowship and the high-performance computing service (PSMN) at the ENS de Lyon.

M.G. acknowledges support by the ERC Advanced Grant STRATUS (ERC-2011-AdG No. 291198).

Notes and references

[a] * *European Centre for Theoretical Studies in Nuclear Physics and Related Areas (ECT*), Bruno Kessler Foundation, and Trento Institute for Fundamental Physics and Applications (INFN-TIFPA), Trento, Italy E-mail: taioli@fbk.eu*

[b] *Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic*

[c] *School of Science and Technology, University of Camerino, Camerino, Italy and INFN, Sezione di Perugia, Italy*

[d] *Dipartimento di Chimica G. Ciamician, University of Bologna, Bologna, Italy*

[e] *Laboratoire de Chimie, UMR 5182 CNRS et Ecole Normale Supérieure de Lyon, 46 allée d'Italie 69364, Lyon Cedex 07, France.*

1 Mukamel S., *Ann. Rev. Phys. Chem*., 2000, **51**, 691.
2 Cho M., *Chem. Rev*., 2008, **108**, 1331
3 Mukherjee P., Kass I., Arkin I., Zanni M., *PNAS,* 2006, **103**, 3528.
4 Cheng Y-C., Fleming G.R., *Annual. Rev. Phys. Chem.,* 2009, **60**, 241
5 Sarovar M., Ishizaki A., Fleming G.R., Whaley K.B., *Nat. Phys.,* 2010, **6**, 462
6 Selig U., Schleussner C.-F., Foerster M., Langhojer F., Nuernberger P., Brixner T., *Opt. Lett.,* 2010, **35**, 4178
7 West B. A., Moran A., M. *J. Phys. Chem. Lett.,* 2012, **3**, 2575
8 Tseng C.-h., Sandor P., Kotur M., Weinacht T. C., Matsika S., *J. Phys. Chem. A*, 2011, **116**, 2654
9 West B. A., Womick J. M., Moran A. M., *J. Phys. Chem., A* 2011, **115**, 8630
10 Nenov A., a Beccara S., Rivalta I., Cerullo G., Mukamel S., Garavelli M., *Chem. Phys. Chem.,* 2014**, 15**, 3282
11 Drescher M. et al., *Nature* 2002, **419**, 803
12 a Beccara S., Fant L., Faccioli, P., 2015, **114**, 098103
13 Taioli S., Simonucci S., Calliari L., Dapor M., *Phys. Rep.,* 2010, **493**, 237
14 Taioli S., Simonucci S., Calliari L., Filippi M., Dapor M., *Phys. Rev. B,* 2009, **79**, 085432
15 Taioli S., Simonucci S., Dapor M., *Comput. Sci. Discov.,* 2009, **2**, 015002
16 Köhler L., Kresse G., *Phys. Rev. B* 2004, **70**, 165405
17 Fano U., *Phys. Rev.,* 1961, **124**, 1866
18 Kresse G.; Hafner J., *Phys. Rev. B,* 1993, **47**, 558
19 Kresse G., Hafner J., *Phys. Rev. B,* 1994, **49**, 14251
20 Kresse G., Furthmuller J., *Comput. Mater. Sci.,* 1996, **6**, 15
21 Alfè D., *Comp. Phys. Comm.,* 1999, **118**, 31
22 Blochl P. E., *Phys. Rev. B,* 1994, **50**, 17953
23 Taioli S., Umari P., De Souza M., *Phys. Stat. Sol.,* 2009, **246**, 2572.
24 Umari P., Petrenko O., Taioli S., De Souza M., *Comm.: J. Chem. Phys.,* 2012, **136**, 181101.