Università degli Studi di Padova

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Biologia

_____

SCUOLA DI DOTTORATO DI RICERCA IN BIOSCIENZE E BIOTECNOLOGIE
INDIRIZZO BIOCHIMICA E BIOFISICA
CICLO XXVII

THE ROLE OF TIME IN BIOLOGICAL SYSTEMS: A COMPUTATIONAL ANALYSIS RANGING FROM MOLECULAR DYNAMICS TO BIOLOGICAL NETWORK SIMULATIONS.

Direttore della Scuola : Ch.mo Prof. GIUSEPPE ZANOTTI
Coordinatore d'indirizzo: Ch.mo Prof. FABIO DI LISA
Supervisore :Ch.mo Prof. SILVIO TOSATTO

Dottorando : ALESSANDRO MASIERO

THE ROLE OF TIME IN BIOLOGICAL SYSTEMS: A COMPUTATIONAL ANALYSIS RANGING FROM MOLECULAR DYNAMICS TO BIOLOGICAL NETWORK SIMULATIONS.

Alessandro Masiero

INDEX

Summary

Time is the main character of this thesis, and it has been used in many ways. Starting from molecular dynamics simulations to biological networks, time has been analysed under different light and roles. The main subjects characterizing this work are Von Hippel-Lindau syndrome and circadian rhythm, although a method for molecular dynamics simulations trajectory analysis is presented at the end. Several works are presented here, analysing different aspects of protein dynamics as well as pathway alterations depending on the time coordinate. The first work analyses the interaction between protein Von Hippel-Lindau (pVHL) and its main interactor, Hipoxya Inducibe Factor 1α by means of molecular dynamics simulations, investigating on a non-conventional proline hydroxylation event. As a result, we obtained that a specific hydrogen bond network rearrangement and improved electrostatic energy for hydroxylated P567 appear to be compatible with an increase in HIF-1α binding affinity. Sequence analysis also confirms P567 to be vastly conserved during evolution, indicating a possible role for this alternative, PHD-3 driven, post translational modification in pVHL–HIF-1α complex formation. The second work dealt with the same main subject, but investigated through biological network simulations, particularly with Petri net models. In this work, we presented a novel manually curated Petri Net (PN) model of the main pVHL functional pathways. The model was built using functional information derived from the literature. It includes all major pVHL functions and is able to credibly reproduce VHL syndrome at the molecular level. The reliability of the PN model also allowed in silico knockout experiments, driven by previous model analysis. Interestingly, PN analysis suggests that the variability of different VHL manifestations is correlated with the concomitant inactivation of different metabolic pathways. In the third work, investigating the structural role of flavin-adenine-dinucleotide (FAD) through molecular dynamics simulations, we analyzed the *Drosophila melanogaster* Cryptochrome crystal structure, elucidating how this large co-factor within the receptor could be crucial for CRY structural stability. The co-factor appears indeed to improve

receptor motility, providing steric hindrance. Moreover, multiple sequence alignments revealed that conserved motifs in the C-terminal tail could be necessary for functional stability. The fourth work focused on the sequence impact on the modern folds. We shuffled the sequences of 10 natural proteins and obtained 40 different and apparently unrelated folds. Our results suggest that shuffled sequences are sufficiently stable and may act as a basis to evolve functional proteins. The common secondary structure of modern proteins is well represented by a small set of permuted sequences, which also show the emergence of intrinsic disorder and aggregation-prone stretches of the polypeptide chain. The last work presented here is a method to quickly analyse molecular dynamics simulations trajectories. The complexity related to their interpretation and analysis is still one of the major challenges for most users. In this work we introduce RING MD, which is able to identify the most important frames (PDB structures) and key residues that cause different conformers transitions, providing a simple interpretation useful for non-expert users. Comparison with the classical analysis of three MD simulations, Ubiquitin, T4 Lysozyme and T4 Glutaredoxin, confirmed RING MD results and effectiveness.

At the end, time should not be considered simply as something entraining the environment, it is what indeed modifies systems and environment. Different systems simply change in different ways, because of different mechanisms, but the main driving force should always be considered time.

Riassunto

Il tempo è soggetto principale di questa tesi, ed è stato utilizzato in molti modi. A partire da simulazioni di dinamica molecolare a reti biologiche, il tempo è stato analizzato in diversi modi e con diversi ruoli. I principali temi che caratterizzano questo lavoro sono la sindrome di Von Hippel-Lindau e il ritmo circadiano, anche se un metodo di analisi di traiettorie di simulazioni di dinamica molecolare è presentato alla fine. Diversi lavori sono presentati qui, analizzando diversi aspetti della dinamica delle proteine, nonché

alterazioni di pathway dipendenti dalla coordinata temporale. Il primo lavoro analizza l'interazione tra proteina Von Hippel-Lindau (pVHL) e il suo interattore principale, Hipoxya Inducibe Factor 1α mediante simulazioni di dinamica molecolare, indagando su un evento di idrossilazione di una prolina non convenzionale. Come risultato, abbiamo ottenuto che uno specifico riarrangiamento di una rete di legami a idrogeno e un incremento di affinità elettrostatica nel sistema contenente la P567 idrossilata sembrano essere compatibili con un aumento di affinità di legame per HIF-1α. Analisi di sequenza confermano anche che P567 è notevolmente conservata durante l'evoluzione, indicando un possibile ruolo per questa modificazione post trasduzionale dipendente dalla PHD-3, nella formazione del complesso pVHL-HIF-1α. Il secondo lavoro affronta lo stesso soggetto principale, ma studiato attraverso simulazioni di reti biologiche, in particolare con reti di Petri. In questo lavoro, abbiamo presentato in un modello curato manualmente di Petri Net (PN) I principali pathway funzionali di pVHL. Il modello è stato costruito utilizzando informazioni funzionali derivate dalla letteratura. Include tutte le funzioni principali pVHL ed è in grado di riprodurre in modo credibile la sindrome VHL a livello molecolare. L'affidabilità del modello PN ha permesso anche esperimenti di knockout *in silico*, guidata da precedenti analisi del modello. È interessante notare che l'analisi della PN suggerisce che la variabilità delle diverse manifestazioni VHL è correlata con l'inattivazione concomitante di diverse vie metaboliche. Nel terzo lavoro, concernente il ruolo strutturale del flavin-adenina-dinucleotide (FAD) attraverso simulazioni di dinamica molecolare, abbiamo analizzato la struttura cristallografica di Criptocromo di *Drosophila melanogaster*, per chiarire come questo co-fattore nei recettore potrebbe essere cruciale per la stabilità strutturale della proteina. Il co-fattore sembra promuovere la motilità del recettore, fornendo un effetto sterico. Inoltre, allineamenti multipli di sequenza hanno rivelato che i motivi conservati nella coda C-terminale potrebbero essere necessari per la stabilità funzionale. Il quarto lavoro concerne l'impatto che diverse sequenze hanno sui fold moderni. Abbiamo mischiato le sequenze di 10 proteine naturali e ottenuto 40 diversi e apparentemente non collegati fold. I nostri risultati suggeriscono che le sequenze ottenute sono sufficientemente stabili e possono fungere da base per evolvere proteine funzionali. La struttura secondaria comune di proteine moderne è ben rappresentata da un piccolo insieme di sequenze permutate che sottolineano anche l'emergere di

disordine intrinseco e tratti di segmenti peptidici pro-aggregazione. L'ultimo lavoro qui presentato è un metodo per analizzare rapidamente traiettorie di simulazioni di dinamica molecolare. La complessità legata alla loro interpretazione e l'analisi sono ancora tra le principali sfide per la maggior parte degli utenti. In questo lavoro introduciamo RING MD, che è in grado di identificare i frame più importanti (strutture PDB) e residui chiave che causano transazioni tra diversi conformeri, fornendo una fonte di semplice interpretazione utile per gli utenti non esperti. Il confronto con l'analisi classica di tre simulazioni, ubiquitina, lisozima e glutaredossina di batteriofago T4, conferma l'efficacia di RING MD. Alla fine, il tempo non dovrebbe essere considerato semplicemente come spettatore passivo, ma effettivamente parte attiva nella modifica di sistemi e dell'ambiente. Diversi sistemi mutano in maniera diversa, a causa di diversi meccanismi, ma la principale forza motrice deve sempre essere considerata il tempo.

1. Introduction

Time is the name given to that dimension through which we are allowed to sort events by chronological order and measure the duration of intervals between them. Whether it had a beginning or not, Time has the absolute value of embedding physical changes, as it contains itself the concept of motion in space. No changes would be admitted without Time, and this mutual relationship, although appearing taken from granted, is nonetheless probably the strongest requirement for the existence of Life. The need for a measure of Time brought it to be part of the seven fundamental measures in the International System of Units, and the majority of the derived units are based on a Time scale. In biology, its strict importance can be found, firstly, in the concept of Life itself. Every organism is born, lives and dies in a Time interval. At a more deep view, we can find infinite possibilities to evaluate its role in biological events. The duration of a human being life is usually measured in years, the duration of a seasonal change in months, the life duration of a butterfly in days, and so on with faster events within smaller intervals. In my work, the concept of Time is recurring in many ways. Undeniably, molecular dynamics simulations are inseparable from Time, but several other roles can be found in my works. The day/night cycles light entrainment of a fly, the impact of evolution, which in turn subtends Time, on protein folds, the steps of a Petri net evolution are some of the examples of the Time impact and presence during my entire Ph.D. experience. My relationship with Time has been characterized by the usage of this dimension "as a tool and not as a couch"(1). The main subjects I have been working on are Von Hippel-Lindau syndrome and circadian rhythms, exploiting state-of-the-art *in silico* methodologies such as molecular dynamics simulations and different types of biological networks.

1.1 Von Hippel-Lindau syndrome

Critical deregulation of cellular pathways, when undergoing without thorough control, often gives rise to a cluster of related diseased commonly termed cancer (2). Particularly, Von Hippel-Lindau (VHL) syndrome is a dominantly inherited familiar cancer syndrome with variable expression and age-dependent penetrance (3). The main character behind this syndrome is the Von Hippel-Lindau protein (pVHL). This protein is the product of the von Hippel-Lindau gene, located in the short arm of the third chromosome and transcribed in both fetal and adult tissues (4). pVHL is a multipurpose adaptor (Fig. 1)(5), engaging in multiple protein-protein interactions and can be found in two isoforms, VHL30 (213 AA) and VHL19 (160 AA), derived from two alternate translational start sites (5). Neither VHL30 nor VHL19 isoforms contain enzymatic domains (5). The best characterized protein Von Hippel-Lindau (pVHL) function is its role as target recognition component of the E3 ubiquitin ligase which targets the α subunits of the hypoxia-inducible factor-1 (HIF-1α) and hypoxia-inducible factor-2 (HIF-2) for degradation in the presence of oxygen. Protein pVHL binds the HIF-1α subunits and the elongin-C (ELO-C) protein, which interact with elongin-B (ELO-B) and cullin-2 forming the VCB complex. The interaction between HIF-1α and pVHL is driven by proline hydroxylation events on this protein, carried out by prolylhydroxylases (PHDs). PHDs are a heterogeneous enzyme family composed of three different proteins (PHD-1 to -3). While PHD-1 and -2 require a specific target sequence (LxxLAP motif) PHD-3 seems to be less sequence specific and able to carry out its hydroxylation activity with the only requirement of a proline residue in the linear motif (6,7). Specific interactions of VHL have been studied and the involvement of VHL in different signaling pathways has been proved, such as microtubule dynamics, cell proliferation, neuronal apoptosis and responses to DNA damage. Several oncosuppressors and oncogenes interacting with pVHL were already identified such as p53 (8) and p14ARF (9). p14ARF regulates p53 degradation binding the MDM2 protein (murine double minute clone 2), which is the main effector of p53 degradation (10). Recently, an interaction between full lenght pVHL and p14ARF was demonstrated, with pVHL30 binding p14ARF and pVHL19 (missing the N-terminal acidic region) lacking this interaction (11). VHL syndrome, the pathological

outcome overlying mutation, truncation or complete deletion events on pVHL, is characterized by cysts and tumors (3,12) such as retinal and central nervous system hemangioblastomas, pheochromocytomas, clear cell renal carcinomas (CCRCs), islet cell tumors of the pancreas, as well as cysts and cystadenomas in the kidney, pancreas, epididymis, and broad ligament (13). Several specific phenotypes characterize the main clinical manifestations. Central nervous system haemangioblastomas are found in 60-80% of the cases, and appear to be one of the most occurring presenting features of the syndrome and jointed with retinal angiomas or haemangioblastomas are the most common clinical symptoms driving towards the syndrome diagnosis (14). One of the most important causes of death linked to VHL syndrome is CCRS (14). Pheochromocytomas are adrenal gland tumors bringing to unbalanced adrenergic activation. Less frequent within the overall VHL population but characterized by high penetrance in elderly patients are pancreatic islet cells tumors (14). VHL syndrome is considered a severe autosomal dominant genetic disease with inheritance of one person over 35,000 (15,16), and being a genetic disorder it follows the Knudson's two hit principle. A copy of the gene is mutated in the germ line, but the other copy still produces a functional protein. Complete pVHL inactivation appears during life, as main consequence of somatic inactivation of the functional copy (17). Mutations occurring in early fetal formation yield to embryonic lethality (18). In this work, I investigated molecular details driving pVHL complex formation, as well as its behaviour within the hypoxia regulation pathway, exploiting bioinformatics and molecular modelling tools.
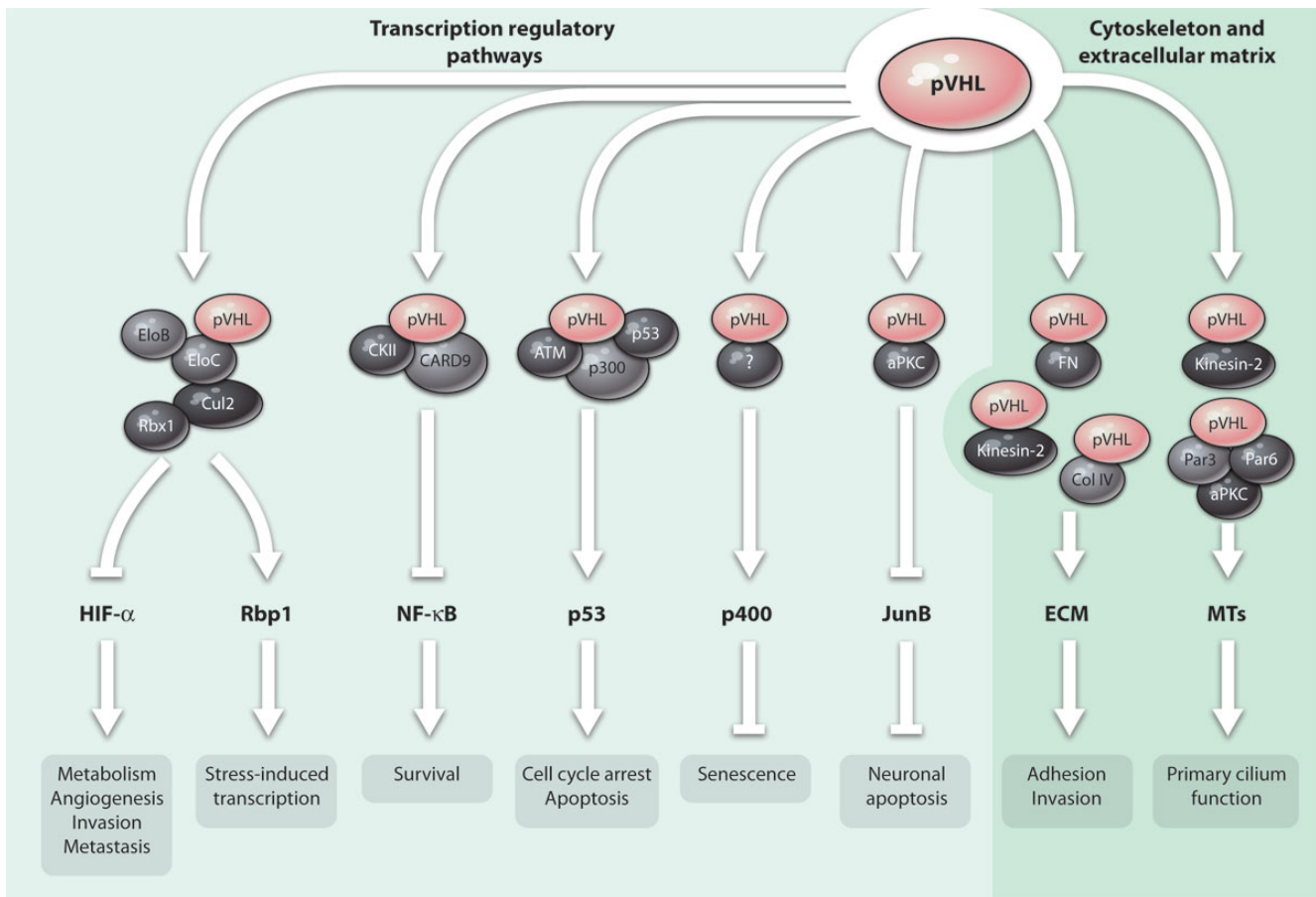
Figure 1: (Frew IJ and Krek W., Sci.Signal., 2008) Summary of the diverse adaptor protein functions of pVHL that regulate both transcription-dependent and transcription-independent cellular responses. pVHL functions as the recognition subunit of the VCB E3 ubiquitin ligase complex composed of Cul2, Elongin B and C (EloB and EloC), and Rbx1 that targets HIF-α and Rbp1. pVHL also interacts with and inhibits aPKC, which may be partly mediated by the VCB complex or may be mediated by a separate complex. As a consequence of the inhibition of aPKC activity, the activity of JunB in antagonizing c-Jun–mediated apoptosis is decreased. pVHL acts as a linker to bring CKII into contact with the substrate CARD9, thereby inhibiting NF-κB signaling. pVHL interacts with p53, ATM, and p300 to nucleate a multimeric complex that leads to p53 activation. pVHL participates in an unidentified protein complex through which pVHL posttranscriptionally stimulates the accumulation of p400 and inhibits cellular senescence. The interaction of pVHL with collagen IV (Col IV) and fibronectin (FN) likely underlies its function in ensuring the correct formation of the ECM. pVHL promotes MT stabilization and associates with the MT motor protein kinesin-2 and the polarity proteins Par3 and Par6, as well as aPKC. Note that with the exception of the VCB E3 ubiquitin ligase complex, the topology and binding domains of the protein-protein interactions in pVHL-associated protein complexes have not been fully characterized. An arrow (↓) represents a stimulatory or activating effect on a protein or cellular response; a blunted arrow (⊥) represents an inhibitory effect; a question mark (?) represents an unknown subunit of a pVHL complex.

1.2 *Drosophila melanogaster* cryptochrome and circadian rhythm entrainment

Light sensing is a crucial component of living organisms, which have evolved as answer to a rhythmic environment characterized by endless light–dark cycles. It is present in all kingdoms of life, and this function is achieved by several different mechanisms. One of these is light sensing by means of a class of proteins termed photoreceptors (19). These proteins are optimally suited for studying fold changes due to light activation (19). The main classes into which these proteins can be divided are six (see Fig. 2), each one characterized by a different photochemical mechanism driving receptor activation. A mechanism of activation clustering three of these six classes exploits a cis/trans transition of different cofactors. The photoreceptors belonging to the first class are Phytochromes, which are activated by a cis/trans transition of the tetrapyrrole chromophore (19,20). The second class photoreceptors are Rhodopsins, which share the activation mechanism of phytochromes but exploit a different cofactor, retinal (19,21). The third class is characterized by Xanthopsins, sharing in turn a cis/trans transition driven activation mechanism provided in this case by coumaric acid and activated by blue light (19,22). The last three classes also share a common cofactor, flavin adenin dinucleotide (FAD). The fourth class, activated by a FAD mediated electron transfer, is gathered under the name of Cryptochromes (CRYs), and are, like Xanthopsins, activated by blue light (19,23). These photoreceptors were first identified in *Arabidopsis thaliana* in 1993 (23). The fifth class has a peculiar activation mechanism which requires the formation of cysteinil adducts, and the photoreceptors belonging to this class are termed Phototropins (19,24). The last class has been found in photophobic responses of *Euglena gracilis* (19). It is called BLUF (sensors for Blue Light Using FAD) and differently from CRYs, which are activated by electron transfer following blue light absorption, the members of this photoreceptor family are believed to be activated by proton transfer following blue light absorption (19,25). In my structural biology work, I focused mainly on Cryptochromes. Indeed, in 2011 Zoltowksi et al published the first crystal structure of the dark state *Drosophila melanogaster* Cryptochrome (dCRY) (26), later updated in 2013 (27) (PDB ID: 4GU5). This photoreceptor is known to share the same tertiary

structure with *Drosophila melanogaster* (6-4) Photolyase (dPL), a DNA repair enzyme reverting purine DNA bases dimerization, but showing a longer C-Terminal Tail (CTT) (28). CRYs, differently from PLs, do not show DNA repair activity in animals (29). The main function of dCRY is the maintenance, in specific subsets of the clock neurons and in the compound eyes of the flies, of day/night cycles sensing light entrainment pathway, and it is carried out by its interaction with TIMELESS (TIM), subsequent to CTT opening brought by FAD reduction upon blue light exposure (30–33). In darkness conditions, PERIOD (PER) binds to TIM to form a dimer. The PER-TIM heterodimer enters the nucleus and inhibits the CLOCK and CYCLE protein interaction (34,35). Light exposure causes dCRY activation and subsequent TIM binding, avoiding its interaction with PER. This interaction, in turn, allows CLOCK and CYCLE interacting with a specific E-box DNA segment promoting the transcription of clock genes (33,36). The absorption by FAD of a photon and the subsequent reduction towards semiquinone state allow CTT to expose plenty of linear motifs, particularly PDZ domains, driving several protein-protein interactions (29).
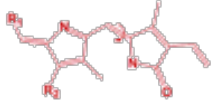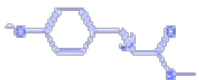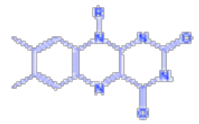
Figure 2: Van der Horst et al. Acc. Chem. Res., 2004. The curved arrow identifies the vinyl bond subject to photoisomerization. In retinal, both the 11,12- and the 13,14-vinyl bonds can undergo isomerization, like in mammalian and in bacterial sensory rhodopsins, respectively (17).

1.2 Protein fold evolution during time

20 is the number of the building blocks of proteins, amino acids. As for a language, composed by 20 letters, the potential number of words and meanings is infinite. The same could be stated for proteins. In fact, what is observed in nature is slightly different. Plenty of protein sequences and structures are conserved, highlighting a relatively small number of folds if compared to the potential infinite number of polypeptides chains that could be obtained from the amino acidic alphabet. Starting from the formation of the primordial broth and the first peptides, proteins and their fold underwent several modification processes rising from function pressure or providing function specialization. Same catalytic activities are found in proteins not sharing the same fold, and vice versa, proteins sharing the same fold are not always sharing the same function. A stable fold is more favourable to be conserved during evolution, and proteins autonomously evolved maintaining the most stable folds. Starting from these assumptions, the question whether the modern folds are an exquisite result of natural selection or rather random sequences slightly improved by evolution arises (37). Considering a prebiotic scenario, it is plausible that modern protein folds are the result of small proto-protein folds. In this work, we investigated through *ab initio* predictions of non-natural proteins on this subject, trying to provide clues on the origins of modern folds.

1.4 Methodologies

1.4.1 Molecular dynamics simulations

In 2013, the Nobel Prize for chemistry has been awarded to the developers of a molecular force field for the simulation of biological systems (38). This framework provides the possibility of investigating the behaviour of molecular structures coming from the Protein Data Bank(39) or from structure prediction methods during time. Indeed, Molecular Dynamics (MD) simulations are the study of successive configurations of the system generated by integration of the Newton's law of motion (40). This law is formed by three rules: (i) a body continues to move in a straight line at constant velocity

unless a force acts upon it, (ii) force equals the rate of change of momentum and (iii) to every action there is an equal and opposite reaction (40). Molecular dynamics simulations were first introduced by Alder and Wainwright in 1957 (41), and during years have been increasingly used, due to the computational power improvement characterizing the last thirty years. This improvement allowed passing from simulations of few femtoseconds to the state-of-the-art micro and also milliseconds long simulations (42,43). In classical MD simulations, atoms are considered as charged spheres owning a mass, and bonds are treated as springs obeying to the Morse potential. The force field contains all charges and masses, bonds, angles, dihedrals and improper torsions parameters. The integrator performs the calculations of Newton's laws based on the force field. Before the simulation starts, the system is placed within a box of water or in a lipidic bilayer, which can be cubic, spherical or octahedral. The used water model for all simulations in this work was TIP3p (44,45). The following step consists in global charge neutralization by means of Na^+ or K^+ positive ions and Cl^- negative ion addition whether the charge to neutralize is negative or positive. Finally, the system is added of a number of the previously mentioned ions to reach physiological concentrations. The solvated system undergoes the equilibration phase, which consists in structure relaxation and incremental heating steps followed by two short integrations first in NVT conditions (constant number of molecules, volume and temperature) and then in NPT (constant number of molecules, pressure and temperature) conditions. Once the system is equilibrated, the production run is carried out. The main integration software packages used in this work are GROMACS (46) and NAMD (47). The former was set to integrate Newton's laws following the leap-frog scheme, while the latter using the Verlet method (40). The leap-frog scheme is a variation of the classic Verlet integration method. The classic Verlet method uses the positions and accelerations at time t, and the positions of the previous step to calculate the new positions in the following timestep as follows:

$$r(t + \delta t) = 2r(t) - r(t - \delta t) + \delta t^2 a(t)$$

being a the acceleration and t the time. The leap-frog algorithm calculates the velocities $v(t + \frac{1}{2}t)$ from the velocities at time $t - \frac{1}{2}t$ and the accelerations at time t. The positions

are deduced by the velocities using the following $r(t + \delta t) = r(t) + \delta t v \left(t + \frac{1}{2} \delta t \right)$ (40). Using these integration software packages, the result is a trajectory, which is then analysed in terms of root mean square deviation (RMSD) and root mean square fluctuation (RMSF). The RMSD is an average distance between a set of atom pairs and is usually stated as follows:

$$RMSD = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \delta_i^2}$$

being δ the distance between N pairs of atoms. The RMSF can be stated as the mean per residue RMSD. These data, joint with visual inspection, allow obtaining molecular information on the behaviour of the analysed system.

1.4.2 Petri net simulation networks

Networks are commonly defined as clusters of objects sharing connections between each other. Several systems have been represented with networks, including biological systems. All pathways occurring to satisfy that fundamental condition of life, which is called homeostasis, can be represented as networks. Mathematical laws usually drive biological interconnections between different interactors. Starting from this assumption, we can build a model of a biological network obeying these laws (48). Petri Nets (PNs) are mathematical representations of distributed systems depicting discrete and concurrent processes in a clear way, immediately highlighting the order of occurrence of several biochemical reactions belonging to a defined pathway (49). They were first introduced by Petri C.A. in 1962 to describe technical systems (50), but the impact of these kinds of representations for biological systems is growing (16,51), given that different kinds of networks are increasingly being used in biology (52). The structure of PNs is simple, but it allows representations of huge and cross-linked biological pathways by means of simple structural objects. These objects are circle/places, squares/transitions, directed arcs and tokens. The nodes of the networks are

constituted by places and transitions, while directed arcs represent the edges (53). Places are the characters of the pathway, such as ions, molecules, proteins or cellular components, while transitions represent chemical or biochemical reactions, molecular modifications or location/state modifications, post translational modifications or synthesis/degradation. Directed arcs drive the way nodes interact each others, starting from the assumption that one places can only link to transitions and vice versa. The last components, tokens, are dots representing the unity inside a place (pre-place or post-place). The move from pre-places to post-places through transitions, representing the possibility of occurrence of a given transition. If the number of tokens satisfies the required pre-conditions, or stoichiometry, thus the transition is allowed to occur and tokens move from pre-places to post-places. In this work, we used Standard PNs, that do not consider the timing variable. Despite this, several types of PNs are commonly used such as Extended PNs, Time PNs, Coloured PNs, and Hybrid PNs just to cite few of them. These several types allow the user to include several different variables in respect to Standard PNs. In order to simplify the final representation of the networks, smaller pathways embedded in the biggest ones can be represented as Macronodes. Macronodes are represented by boxed nodes, and can be arbitrary nested in multiple layers to keep the represented network clear.

1.5 Aim of the thesis

The possibility offered by the modern computational power to perform fast calculations on huge systems provided a useful framework to investigate the time impact on several biological aspects. The usage of modern techniques such as molecular dynamics simulations of biological systems or Petri net simulations of biochemical pathways allowed me to investigate, sometimes provide explanation, and formulate hypothesis towards several biological and biochemical issues. The growing impact and reliability that bioinformatics and molecular modelling tools are having nowadays were at the basis of my Ph.D. and scientific experience so far. In spite of providing some merely theoretical answers, the capability of *in silico* predictions of simulating real systems is

gaining light. With this work, I provided some insights on several time related issues, such as the cruel impact of time in cancer related diseases, as well as the more lenient impact it has on protein fold evolution, and on day/night cycles. Both cancer and circadian rhythm entrainment are important scientific topics, involving a huge community of scientist investigating on them.

The aim of this work was to use modern computational approaches to provide some clues on both pure biochemical and more applied biomedical issues, as answer to the previously mentioned resources availability. Under this light, each hypothesis for each system was investigated by several and different computational approaches, ranging from sequence analysis to the simulation of peculiar molecular interactions, passing through a huger analysis of an entire pathway. The main variable, which was also chosen to be part of the title of the present work, is time. The latter allowed seeing the changes occurring during the several dynamic processes I worked on.

1.6 Outline

The following chapters represent all the tasks I have been working during my Ph.D. experience. Chapters 2 to 5 are published works, particularly:

Chapter 2 is published in *Minervini G, Masiero A, Moro S, Tosatto SCE. In silico investigation of PHD-3 specific HIF1-α proline 567 hydroxylation: A new player in the VHL/HIF-1α interaction pathway? FEBS Lett. 2013 Sep 17;587(18):2996–3001.*

Chapter 3 is published in *Minervini G, Panizzoni E, Giollo M, Masiero A, Ferrari C, Tosatto SCE. Design and Analysis of a Petri Net Model of the Von Hippel-Lindau (VHL) Tumor Suppressor Interaction Network. PloS One. 2014;9(6):e96986.*

Chapter 4 is published in *Masiero A, Aufiero S, Minervini G, Moro S, Costa R, Tosatto SCE. Evaluation of the steric impact of flavin adenine dinucleotide in Drosophila melanogaster cryptochrome function. Biochem Biophys Res Commun. 2014 Aug 8;450(4):1606–11.*

Chapter 5 is published in *Minervini G, Masiero A, Potenza E, Tosatto SCE. Structural protein reorganization and fold emergence investigated through amino acid sequence permutations. Amino Acids. 2015 Jan;47(1):147–52.*

Chapter 6 represents the last task I performed during my Ph.D. experience, and is still

unpublished, while Chapter 8 contains all supplementary information attached to published papers, to facilitate the reader in finding what is mentioned in the main texts.

2. In silico investigation of PHD-3 specific HIF1-α proline 567 hydroxylation: a new player in the VHL/HIF-1α interaction pathway?

This chapter has been published in (54) Minervini G, Masiero A, Moro S, Tosatto SCE. In silico investigation of PHD-3 specific HIF1-α proline 567 hydroxylation: A new player in the VHL/HIF-1α interaction pathway? FEBS Lett. 2013 Sep 17;587(18):2996–3001.

2.1 Summary

Hypoxia inducible factor 1α (HIF-1α) regulates oxygen homeostasis in the cell through a sensing mechanism involving its hydroxylation and binding to the von Hippel-Lindau (VHL) tumor suppressor. This mechanism is mediated through hydroxylation of HIF-1α proline 564, although in vitro tests have previously shown an alternative hydroxylation at proline 567 by PHD-3. Here, molecular dynamics simulations were used to investigate the structural effect of this alternative hydroxylation. A specific hydrogen bond network rearrangement and improved electrostatic energy for hydroxylated P567 are compatible with an increase in HIF-1α binding affinity. Sequence analysis also confirms P567 to be vastly conserved during evolution, indicating a possible role for this alternative, PHD-3 driven, post translational modification in pVHL-HIF-1α complex formation. The analyzed systems were four: (i) the complex with hydroxylation on P564, (ii) the complex with hydroxylation on P567, (iii) the complex with hydroxylation on both P564 and P567 and (iv) the complex with no hydroxylation. All simulations were carried out with Gromacs, using the CHARMM 27 force field. HYP567 was generated keeping the same stereoisomer as HYP564, and both 4-hydroxyl moieties were placed in the

(R) conformation. Here, we tested the structural compatibility of P567 hydroxylation through MD simulations of the effect of this non-conventional hydroxylation on the interaction interface between pVHL and the HIF-1α linear motif. Our results suggest that a previously not described reorganization of hydrogen bond network between pVHL and HIF-1α appears around the HIF-1α P567 linear motif when hydroxylated. HYP in position 567 is able to promote a new hydrogen bond network localized around HYP567, involving the pVHL H110, Q73, R108 and HIF-1α D569 residues. My contribution to this work ranged from the calculation and implementation of the force field parameters for HYP, being it a nonstandard residue and thus not contained in the force field, ending with the analysis of the trajectory and the electrostatic analysis to obtain the provided results.

2.2 Abstract

Hypoxia inducible factor 1α (HIF-1α) regulates oxygen homeostasis in the cell through a sensing mechanism involving its hydroxylation and binding to the von Hippel–Lindau (VHL) tumor suppressor. This mechanism is mediated through hydroxylation of HIF-1 α proline 564, although in vitro tests have previously shown an alternative hydroxylation at proline 567 by PHD-3. Here, molecular dynamics simulations were used to investigate the structural effect of this alternative hydroxylation. A specific hydrogen bond network rearrangement and improved electrostatic energy for hydroxylated P567 are compatible with an increase in HIF-1α binding affinity. Sequence analysis also confirms P567 to be vastly conserved during evolution, indicating a possible role for this alternative, PHD-3 driven, post translational modification in pVHL–HIF-1α complex formation.

2.3 Introduction

Hypoxia is a pathological condition commonly present in tissue tumor growth. Hypoxia inducible factor 1α (HIF-1α) is a key transcription factor of oxygen homeostasis regulation in the cell. HIF-1α directly regulates the expression of over 40 important target genes such as vascular endothelial growth factor (VEGF), erythropoietin, glycolytic enzymes, and glucose transporters (13). The HIF-1α concentration is regulated through ubiquitin-mediated proteolysis, governed by the activity of the von Hippel–Lindau (VHL) tumor suppressor protein (pVHL) (55). VHL is the key player in VHL syndrome, a dominantly inherited familial cancer syndrome with variable expression and an age-dependent penetrance. It is characterized by a predisposition to develop retinal and central nervous system hemangioblastomas, pheochromocytomas, clear cell renal carcinomas, islet cell tumors of the pancreas, and endolynphatic sac tumors, as well as cysts and cystadenomas in the kidney, pancreas, epididymis, and broad ligament (3,12). The predisposition to develop this variety of tumors is linked to germline inactivation of VHL. Development of the pathology in VHL disease occurs subsequently to somatic inactivation of the remaining wildtype allele in a susceptible cell (5). Importantly, sporadic forms of the same tumor types that are common in familial VHL disease display bi-allelic somatic VHL inactivation through a variety of mechanisms, including mutations and hypermethylation. For example, inactivation of VHL has been demonstrated in 70–80% of all sporadic clear cell renal cell carcinomas (57). pVHL is a multifunctional adapter that interacts with numerous proteins such as ElonginB/C- Cullin2, to form the well characterized VBC degradation complex (58,59), and cytoplasmic microtubules during mitosis (60). Previous computational work on VHL has mainly concentrated on analyzing the many known pVHL interactions (61) and explaining possible effects of known mutations either empirically (62,63) or with molecular dynamics simulations (64–66). VBC mediated degradation of HIF-1α under normal oxygen conditions is
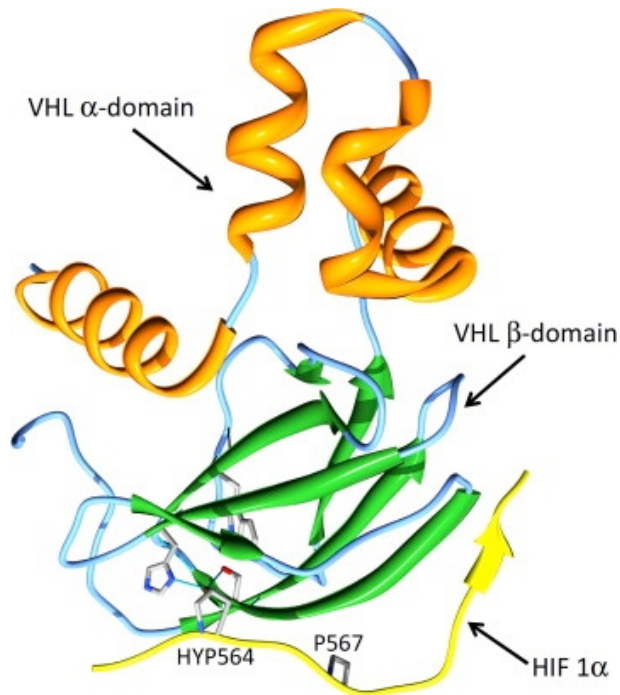
Figure 3: Overview of the pVHL/HIF-1α complex. Cartoon representation of the crystallographic structure of pVHL in complex with HIF-1α (PDB identifier 1LM8). The pVHL α-domain is shown in orange color, while the pVHL β-domain is in green and HIF-1α peptide in yellow. Hydrogen bonds are shown as thin blue lines. The HYP564 and P567 side chains are shown in sticks and labeled.

promoted by oxygen-dependent hydroxylation of P564 and P402 residues operated by members of the HIF-prolyl hydroxylase family (PHDs) (67,68). A crystal structure of the VBC/HIF-1α complex (1.8 Å) including the HIF-1α recognition mechanism operated by pVHL was solved in 2002 (69). The protein–protein interaction is mediated by a highly conserved hydrogen bonds network between pVHL residues S111, H115, W117 and the hydroxylated HIF-1α P564. Under pathological conditions, such as tumor growth, lack of oxygen inhibits normal PHD activity and results in reduced HIF-1α degradation (70).

Loss of HIF-1α degradation then promotes transcriptional activation of numerous genes, resulting in the commonly observed hypervascularized tumors and cysts observed in VHL patients (71). The post translational modification introduced by PHD assumes a crucial role in HIF-1α balance and regulation. The HIF-prolyl hydroxylases are a heterogeneous enzyme family composed of three different

proteins termed PHD-1 to -3. PHD-3, in particular, appears to be a functionally specialized paralog of the first two enzymes. Smaller, and with a markedly different cellular localization, it was recently demonstrated to also have different substrate specificity, at least in vitro (72). While PHD-1 and -2 require a specific target sequence, known as LxxLAP motif (67), PHD-3 seems to be less sequence specific and able to carry out its hydroxylation activity with the only requirement of a proline residue in the linear motif. In 2007, Fedulova and co-workers demonstrated that PHD-3 hydroxylates P567 in HIF-1α (6), which is not LxxLAP motif compliant. In another work, the same residue was also reported to increase the binding affinity of PHD-3(73)). Both observations suggest the possible existence of an alternative HIF-1α regulation pathway mediated by PHD-3. Here, we used molecular dynamics to investigate the structural effect on the pVHL/HIF-1α binding interface when P567 is hydroxylated. Sequence analysis confirms P567 to be broadly conserved during evolution, further strengthening the idea of an evolutionarily conserved functional mechanism.

2.4 Methods

2.4.1 Molecular dynamics simulations

The crystal structure of pVHL solved at 1.8 Å resolution (69)(PDB identifier 1LM8) was used as a starting model, considering chains V (pVHL) and H (HIF-1α peptide). Chains B and C, corresponding to Elongin B and C respectively, were excluded from the simulations as they are distant from the pVHL/HIF-1α interaction site with 40 and 55 Å, respectively and should not have a direct effect on the binding interface interactions of interest for the present work. The pVHL crystal structure also lacks 6 N-terminal and 4 C-terminal residues, which are also not involved in HIF-1α interaction. The protonation state of the protein was adjusted to mimic a pH value of 7.0. All lysine residues were positively charged and histidine residues were

modeled as neutral by protonating the Nε2 atom. In particular, H115 was modeled as neutral by protonating the Nd1 atom as reported in the crystal structure (69). Aspartate and glutamate residues were considered fully deprotonated. All simulations were carried out with Gromacs (46), using the CHARMM 27 (38) force field. Hydrogen atoms were added to the system by means of the Gromacs pdb2gmx routine. Each run was constituted by a minimization step, 100 ps of NVT (constant number of molecules, volume and temperature) simulation, 100 ps of NPT (constant number of molecules, pressure and temperature) simulation and 50 ns of classical molecular dynamics (MD) simulation. The minimization used a steepest descent algorithm. NVT simulations were performed with no pressure coupling. Temperature was coupled with a modified Berendsen thermostat, to obtain a Maxwell distribution of energies. Temperature was kept at 300 K. NPT simulations were performed with a Parinello-Rahman barostat, at a temperature of 300 K, and the pressure was kept at 1 atm. The solvent box was generated covering a distance of 10 Å from the farthest point of the protein boundaries on each of the three dimensions. The generated box was constituted by $\sim 1.8 \times 10^4$ TIP3p water molecules, and sized 82x82x82 Å^3. As HYP is not a standard amino acid, its parameters are not included in standard molecular modeling force fields. A proper set of values, compatible with the CHARMM 27 force field was generated and used for the simulation as follows. The CHARMM force field includes four different contributions, namely bond stretching, angle bending, bond rotation and non-bonded interactions. For the first three we used CHARMM standard values. To obtain partial atomic charges, HYP was examined with the MOPAC AM1 (URL: http://OpenMOPAC.net/) ab initio method. The charges were almost all comparable, with small differences in the CD atom (see Supplementary material), which according to our calculations appears to be slightly positively charged, as it is bound to a peptide nitrogen atom, in agreement with previous AMBER calculations (74). The charges were implemented in CHARMM27 force field, as previously done for collagen peptides (27,28). The resulting values are shown in Supplementary Fig.

1. HYP567 was generated keeping the same stereoisomer as HYP564, and both 4-hydroxyl moieties were placed in the (R) conformation. The total system charge was neutralized by one chloride ion. Default CHARMM parameters were used for ions in bulk solution. All systems were caged into periodic boundary condition cells and Particle Mesh Ewald (PME) grids. All MD runs are 50 ns long and performed on a x86 Linux cluster. The average occupancy was 32 cores per run. To confirm the simulation results shown, two different MD runs were used for each analyzed system. The stereochemical quality of the system at frame 0 was confirmed using PROCHECK(77) and TAP (78).

2.4.2 Electrostatic analysis

Electrostatic interaction energies were derived using APBS (79), which calculates them by splitting the complex into two separate interactors. The transfer free energies are determined from a homogeneous dielectric environment to an inhomogeneous dielectric environment with different bulk and solvent dielectric constants. The difference between the energetic state of the divided interactors with inhomogeneous dielectrics and the complex with inhomogeneous dielectrics represents the binding energy estimate. The APBS calculations were carried out on the last frame of each run.

2.4.3 Sequence analysis

A multiple sequence alignment of the HIF-1α sequence surrounding the hydroxylation site was extracted from Pfam(80)) family PF11413 to map the evolutionary conservation of residue P567. All Metazoan sequences, covering 327 sequences and 100 species were collected (accessed on 25th June 2013). A sequence logo was built with (81) from this data and overlaid with the experimental HIF-1α peptide structure from PDB identifier 1LM8, chain H (69).

Intrinsic disorder for the HIF-1α sequence was predicted with CSpritz (82).

2.5 Results

2.5.1 Molecular dynamics simulations

The pVHL protein is composed of two main domains, termed a- and β-domain, respectively. The α-domain is known to be partially disordered when not involved in interactions with other proteins, while the β-domain is definitely more stable due to the presence of long β-strands (see Fig. 3). To investigate the effect induced by HIF-1α P567 hydroxylation, a 50 ns MD simulation was performed on the pVHL/HIF-1α complex. Four different conditions were postulated and analyzed: hydroxylation of P564 (HYP564), hydroxylation of P567 (HYP567), hydroxylation of both P564 and P567 and the pVHL/HIF-1α complex without post-translational modifications. The unmodified complex was also used as baseline to configure the simulation environment. The variation of the weighted root-mean-square deviations (RMSD) was monitored to investigate the stability of pVHL/HIF-1α interaction. RMSD plots obtained for the four analyzed systems are shown in Fig. 4. The steady RMSD observed for the backbone atoms indicates that the four systems remain stable for the entire simulation time, with the main variations limited to the pVHL a-domain. This result is consistent with the literature (65) and could be considered a reliability check of the MD simulation parameters used. The pVHL/HIF-1α system with no hydroxyproline residues shows a modest but linear RMSD increase during time, suggesting possible major variations over longer simulation times. Analysis of the last MD frame reveals the complex to be compatible with the previously observed crystal structure, revealing small pVHL movements limited to the a-domain. Binding between HIF and pVHL is correctly simulated and consistent with a pseudo β-sheet connection of the HIF linear motif (residues Y565, I566, F572, Q573, L574) and the fourth pVHL β-strand (H110, G106, T105, G104) as previously described (69). The internal HIF-1α motif region

from P566 to D571 shows a stable coil-shaped organization with hydrogen bonds remaining stable during the entire simulation.
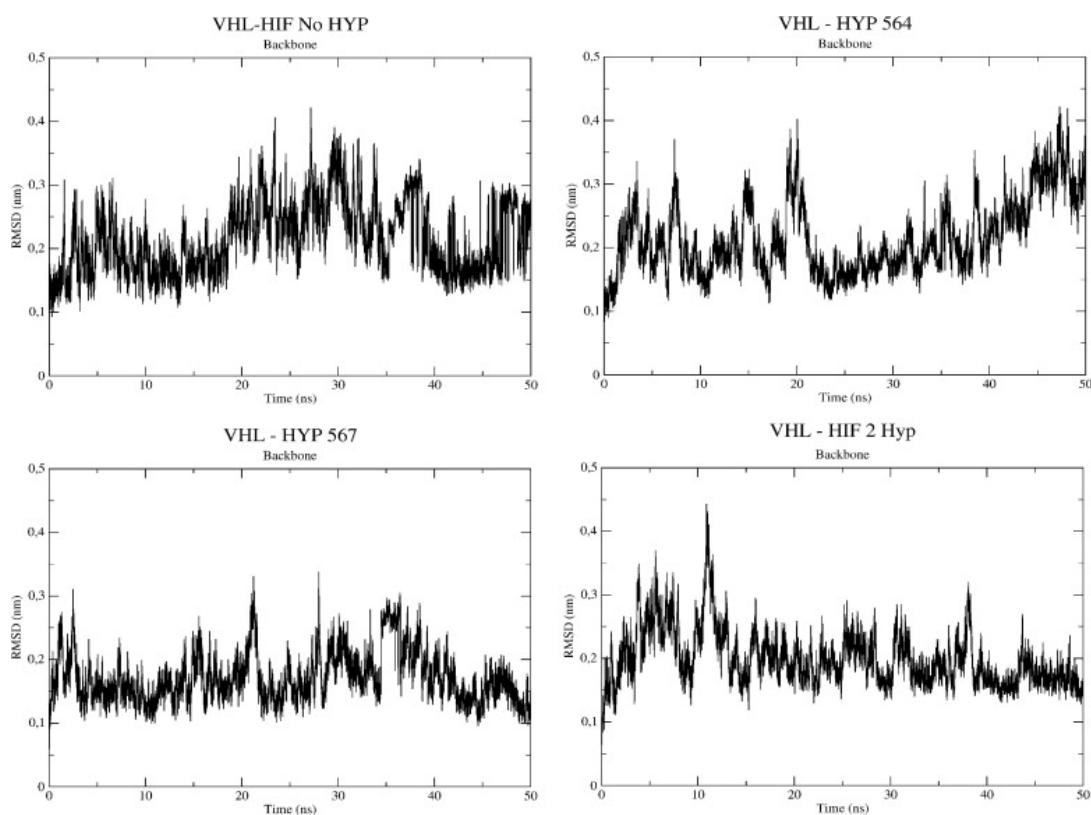
Figure 4: Weighted RMSD fluctuations of the pVHL/HIF-1α complex backbone over time. The four panels show the results without hydroxyproline (VHL-HIF No HYP), with HYP564 (VHL-HYP 564), with HYP567 (VHL-HYP 567) and with double hydroxylation at HYP564 and HYP567 respectively (VHL-HIF 2 Hyp). The RMSD variation was calculated from their initial position during 50 ns of MD simulation.

Similar results for this region were obtained for all tested systems. pVHL/HIF-1α with HYP564 shows smaller fluctuations in the β-domain, probably due to the stabilizing effect, induced by interactions between pVHL S111, H115, W117 and HYP564 (see Fig. 5). Bigger but not relevant variations (RMSD <0.5 Å) were observed for the a-domain, probably related to the absence of a stabilizing effect usually induced by the ElonginB/C-Cullin2 proteins which were not included in our simulations. The RMSD plot clearly shows how the overall system initially remains

stable with little fluctuations. After 40 ns of simulation the RMSD value appears to follow a growing trend (Fig. 4). pVHL/HIF-1α with HYP567 shows less fluctuations, with an estimated average RMSD value around 0.2 Å. Here, we tested if P567 hydroxylation could play a role in pVHL/HIF-1α complex formation. HYP in position 567 is able to promote a new hydrogen bond network localized around HYP567, shown in Fig. 3, involving the pVHL H110, Q73, R108 and HIF-1α D569 residues. This appears to confirm the observation that PHD-3 is able to hydroxylate a proline residue not strictly compliant with the LxxLAP motif. In the double hydroxylated pVHL/ HIF-1α system, the RMSD plot shows an intermediate pattern. For the first 12 ns, the system seems to replicate the previously observed behavior of the non-HYP564, while it and gets closer to the HYP567 behavior for the remaining 38 ns of simulation. A significant spike is visible at 11 ns (Fig. 4) and MD frame analysis reveals a reorganization event of the hydrogen bond network around HYP564 and HYP567. Spikes of similar magnitudes were also noted in RMSD plots for other simulations, but in these cases were related to rearrangement of single side chains involved in the VHL–HIF interaction. In the double hydroxylated system presence of the second hydroxyproline seems to promote an increased distance between HYP564 and pVHL S111 from 2.86 to 4.34 Å, with a consequent partial interruption of the hydrogen bond network. Although a reorientation event of HYP564 and H115 with respect to S111 is visible, the interaction between these residues remains stable during the simulation. The rearrangement of connections seems to be due to the rigidity imposed by the new hydrogen bond network localized around HYP567 and shown in Fig 5. Our simulation results suggest pVHL residues H110, Q73, R108 and HIF-1α D569 to be important to establish the hydrogen bond network around HYP567. Q73 in particular seems to play a key role in establishing this network, although further experimental validation is necessary to confirm its role. Analysis of the pVHL – HIF-1α HYP567 complex trajectory suggests that the P567 modification alone is sufficient to promote the organization of a well structured hydrogen bond network

around this residue. Connections with the pVHL H110, Q73, R108 and the HIF-1α D569 residues are apparently formed due to an increased distance between the coil-shaped region and residues on the fourth pVHL β-strand (data not shown). Based on MD simulation results, residue P567 seems important in the complex formation. To investigate the role of single residues in VHL–HIF-1α interaction, a root-mean-square fluctuation (RMSF) analysis was performed, see Supplementary material.
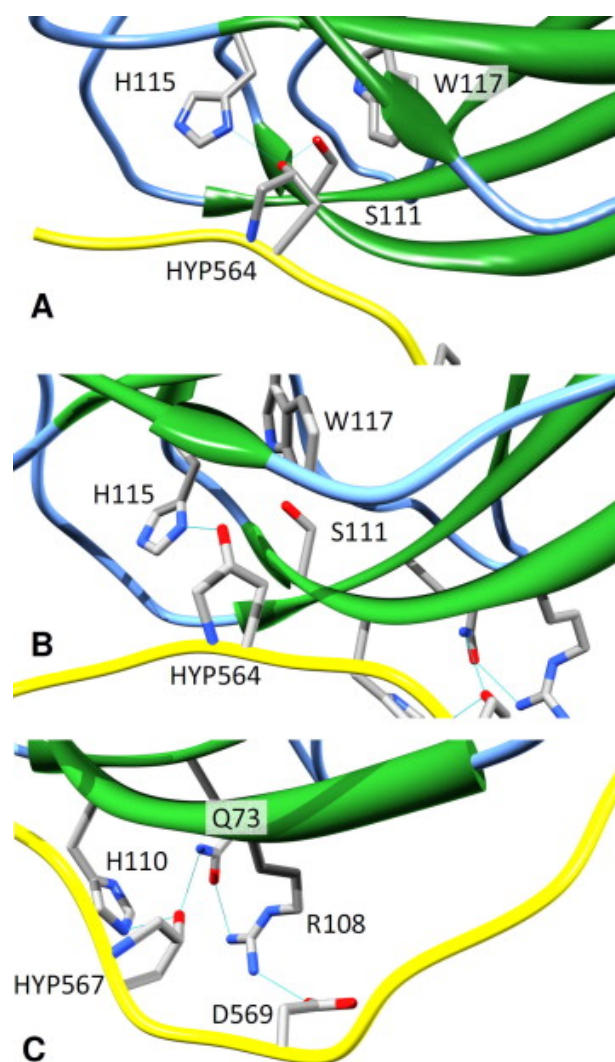
Figure 5: Hydrogen bond network around the hydroxylated HIF-1α residues. A close-up of the pVHL/HIF-1α complex is shown under different conditions in analogy to Fig. 1. Side chains of involved in the hydrogen bond network of the two hydroxylated prolines are shown as sticks, with thin blue lines representing hydrogen bonds. The three panels are organized as follows: (A) Crystal structure with HYP564. (B) Hydrogen bond network reorganization around residue HYP564 promoted by hydroxylation of PRO567. (C) The putative new hydrogen bond network around HYP567. Notice how hydroxylation of P567 creates novel hydrogen bonds with H110 and Q73, while tilting the positions of S111 and W117.

2.5.2 Control MD simulations

In order to check the pVHL/HIF-1α complex stability two control MD simulations were performed. The first one, derived from sequence analysis, containing the

28

P567S mutant, the second one consisting in P564L and P567L mutants. The MD simulations suggest no destabilization of the pVHL/HIF-1α complex with the P567S substitution, consistent with its presence in a subset of HIF sequences. The P564L and P567L double mutant instead presented a significant destabilization leading to gradual dissociation of the HIF-1α peptide from pVHL. A more detailed analysis, including RMSF, can be found in Supplementary material.

2.5.3 HIF1-α motif sequence analysis
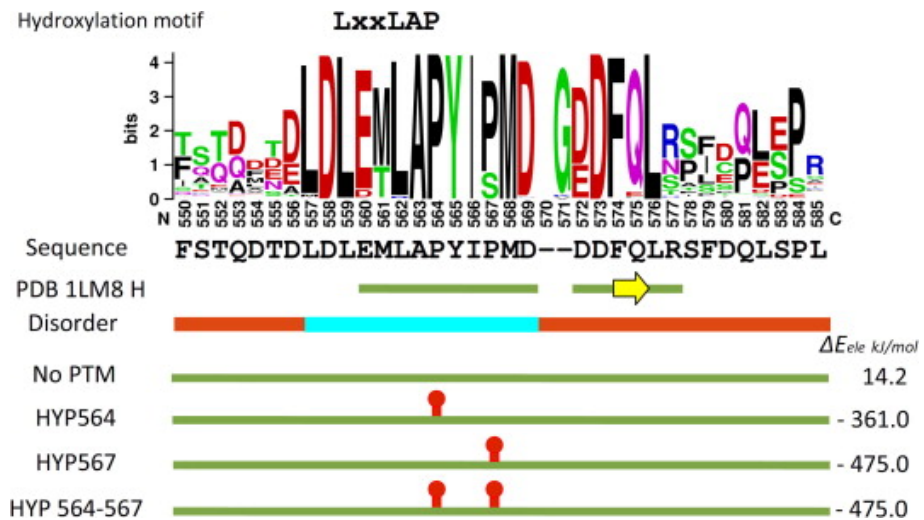
Figure 6: Overview of the HIF-1α peptide features. The HIF-1α sequence logo surrounding the hydroxylation site is shown with the canonical hydroxylation motif above and the human HIF-1α sequence, features from the crystallographic HIF-1α structure (PDB identifier 1lm8) and predicted disorder below. The secondary structure is shown as a green line for coil and yellow arrow for β-strand. Predicted disorder is shown as a red line and predicted structure in light blue. The bottom part shows a schematic representation of the simulated peptides, with red markers depicting the presence of hydroxylated prolines of the two modeled sites (P564 and P567). The electrostatic component of ΔG calculated with APBS is shown to the right of each simulated peptide.

To reinforce this observation, we made a conservation analysis of the HIF-1α linear motif residues among homologous proteins by selecting all Metazoan HIF sequences in Pfam (see Fig. 6). The sequence logo shows how the HIF-1α linear motif is located in an ordered segment with strong sequence conservation, flanked by charged residues on both sides, inside an otherwise intrinsically disordered region. Such an arrangement is generally indicative of a functional linear motif

(83). In particular, P567 to be conserved in 76% of sequences included in the alignment. The only substitution is serine (24% of sequences), a residue hypothetically compatible with the hydrogen bond network observed around P567 (see Fig. 5). To better understand the role of a serine 567 substitution, we analyzed the sequences presenting the substitution. The mutation seems to be present only in HIF-3α orthologs and a less characterized subgroup including sweet water fish of the Cyprini-formes order. HIF-3α has a different expression pattern from HIF-1α and -2α (84). Sweet water ecosystems like shallow lakes or slow rivers during summer are subject to severe oxygen reduction connected with algal bloom phenomena (85). The existence of an adaptive hypoxia tolerance related to environmental conditions was recently demonstrated for a marine species, the common sole (*Solea solea*) (86). In other words, functional specialization among HIF paralogs and seasonal fluctuations of oxygen concentration may have influenced the evolution of an alternative regulation of the hypoxia response pathway including the S567 mutation.

2.5.4 In silico electrostatic binding energy calculation

In order to confirm our results with experimental data, putative electrostatic DG values were calculated with APBS (79) (see Fig. 6), which evaluates energetic states calculating the difference between the solvated unbound interactors and the solvated complex. The values obtained were compared with previously reported experimental values, where the complex containing HYP564 was reported to have DG = -9.1 ± 0.05 kcal/mol (-38.07 kJ/mol), while no interactions could be detected by the authors on the non-hydroxylated HYP564 complex (69). According to APBS, the electrostatic DG contribution of HYP564 to the complex is -371 kJ/ mol, which is consistent with the reported DG trend. Repulsive forces seem to appear in the tested non-hydroxylated complex, with a positive electrostatic contribution of 14.2 kJ/mol, consistent with (69). Based upon these consistent results, electrostatic

contributions were calculated with APBS for all complexes. The resulting putative electrostatic contribution of the HYP564/HYP567 complex was -457.5 kJ/mol. The HYP567 complex showed a very similar putative value of -457.3 kJ/mol. Given the in silico nature of the results, we cannot exclude other interpretations or conclusions except that a set of residues present on pVHL is compatible with a post translational modification of both P564 and P567 residues.

2.6 Discussion

In this work we presented results obtained from MD simulations of the pVHL/HIF-1α complex. Four different hydroxylation patterns (and two controls) were simulated for the proline residues present on a small linear motif of HIF-1α known to interact with pVHL. The simulation started from an observation of Fedulova et al. (6) that PHD-3, a member of human PHD family, hydroxylates the P567 residue not included in the canonical LxxLAP motif usually required by PHD enzymes. PHD-3 differs from other PHD family members by size, cellular localization and target specificity (71). Here, we tested the structural compatibility of P567 hydroxylation through MD simulations of the effect of this non-conventional hydroxylation on the interaction interface between pVHL and the HIF-1α linear motif. Our results suggest that a previously not described reorganization of hydrogen bond network between pVHL and HIF-1α appears around the HIF-1α P567 linear motif when hydroxylated. Recently, Jaakkola and coworkers (87) demonstrated involvement of PHD-3 in apoptotic events under normoxia, apparently related to an over-saturation of the proteosomal degradation system connected with massive protein aggregation. They also demonstrated a massive transcription induction of PHD-3 mRNA promoted by HIF-1α (73). Again, their work demonstrates how PHD-3 activity is maximal when oxygen concentration is restored after strong hypoxia events, probably co-occurring with PHD-1 and -2 reactivation. Our results suggest that the reorganization of the hydrogen bond network also appears when both

HYP564 and HYP567 are present. This scenario is consistent with two distinct post translational modifications due to the subsequent activity of two different PHD enzymes. On the other hand, due to the massive expression of PHD-3 during hypoxia, it is also possible to imagine that double hydroxylation is the result of a higher intracellular concentration of PHD-3. The latter is also known to be actively expressed, under normal conditions, only in heart and brain tissues (72,87,88). Considering our results in context of this evidence, it is easy to speculate that they could indicate an alternative regulation pathway trigged by the PHD-3 enzyme. In other words, we can imagine that hydroxylation of P567 could act as a reinforcement of the ubiquitin–proteasome degradation pathway evolving in highly specialized tissues, which are very sensible to oxygen variation. While this evidence is suggestive, we cannot assume that the obtained computational results are exclusively connected with PHD-3 activity and more complex scenarios where different PHDs act in concert cannot be excluded. Furthermore, data presented in this work suffers from limits and approximations inherent in computational techniques. Despite the apparently good results obtained with 50 ns MD simulations, we cannot exclude complex dissociation on longer timescales. Performing micro or millisecond simulations may provide different outcomes, but is still computationally too expensive to be feasible. Finally, understanding the biological role and relevance of the results presented on the etiopathogenesis of VHL disease will require experimental confirmation, as it is beyond the possibility of the in silico techniques used.

3. Design and analysis of a Petri net model of the Von Hippel-Lindau (VHL) tumor suppressor interaction network.

3.1 Summary

The best-characterized function of pVHL is the ubiquitination dependent degradation of Hypoxia Inducible Factor (HIF) via the proteasome. It is also involved in several cellular pathways acting as a molecular hub and interacting with more than 200 different proteins. Molecular details of pVHL plasticity remain in large part unknown. Here, we presented a manually curated Petri Net (PN) model of the main pVHL functional pathways. The model was built using functional information derived from the literature. It includes all major pVHL functions and is able to credibly reproduce VHL syndrome at the molecular level. The reliability of the PN model also allowed in silico knockout experiments. The network was designed in the Snoopy PN framework. We chose Snoopy to facilitate future extensions of the VHL pathway presented here. Both Charlie and PInA analyzers were used for PN analysis and validation. Further, in silico knock out experiments were used to test the biological reliability of the model. Structural model validation was made by analysis of the T-invariants to demonstrate whether the system was covered by T-invariants and to confirm the biological meaning of each invariant. The computed invariants were grouped in Maximal Common Transition Sets (MCTS) and Clusters, the former based on occurrence of specific sets of transition inside the various T-invariants, and the latter based on similarities between T-

invariants. The validation was made by selectively deleting tokens inside the model, imitating possible biological disruptions such as disease-causing mutations. We started from the transcription activity of HIF due to its regulation is the most studied pVHL function. Our model, as expected from literature data, shows that HIF-1α enters the nucleus when not degraded by pVHL. It subsequently binds HIF-1β to form the HIF heterocomplex, which interacts with DNA. The model simulates the increased affinity of HIF towards DNA. The network itself was the first result we obtained, and we made it publicly available for further implementation. The network proved to be alive, covered in T-invariants and usable for the analysis. The in silico knock out experiments provided the following results:

(i) pVHL knock out. Degradation of HIF-1α is not completely depleted due to presence of both p53- and GSK3β-dependent alternative degradation pathways. All other processes usually inhibited by pVHL take place in an uncontrolled way, including creation of VEGF via Sp1 transcription activity and increased matrix regulation due to lack of fibronectin crosslinking. Hur resulted constantly activated and nur77 can stimulate synthesis of Proopiomelanocortin, precursor for the Adrenocorticotropic hormone. Card9 increases release of tumor necrosis factor, and NF-kB when not inhibited by pVHL. Instead, Jade1 is unable to survive long enough to inhibit β−catenin, generating a proliferation signal with Wnt. Lactic acid is also not produced due to LDH enzyme production being HIF-1α transcription activity dependent.

(ii) HIF-1α knock out. VEGF is still created thanks to Sp1, thus oxygen is still generated even if in lower proportion. If HIF-1α and Sp1 are both knocked out at the same time, oxygen is quickly consumed and the metabolism is soon unable to proceed. Lactic acid is not produced due to LDH enzyme production being HIF-1α transcription activity dependent. Glycolysis and glycogen are produced normally and the metabolism is not inhibited by PyrDH negative regulation and lactic acid formation. Since pVHL is present, other tumor suppressor activities are enabled, except for proteasomal degradation of HIF-1α due to the substrate being non-

existent. My role in this work passed from the validation of protein-protein interactions to the interpretation of the output provided by the analysis.

3.2 Abstract

Von Hippel-Lindau (VHL) syndrome is a hereditary condition predisposing to the development of different cancer forms, related to germline inactivation of the homonymous tumor suppressor pVHL. The best characterized function of pVHL is the ubiquitination dependent degradation of Hypoxia Inducible Factor (HIF) via the proteasome. It is also involved in several cellular pathways acting as a molecular hub and interacting with more than 200 different proteins. Molecular details of pVHL plasticity remain in large part unknown. Here, we present a novel manually curated Petri Net (PN) model of the main pVHL functional pathways. The model was built using functional information derived from the literature. It includes all major pVHL functions and is able to credibly reproduce VHL syndrome at the molecular level. The reliability of the PN model also allowed in silico knockout experiments, driven by previous model analysis. Interestingly, PN analysis suggests that the variability of different VHL manifestations is correlated with the concomitant inactivation of different metabolic pathways.

3.3 Introduction

Pathological deregulation of cellular pathways often results in a family of complex and correlated diseases commonly termed cancer (2). Cancer is a multi factorial disease where different causes contribute to its development. Several computational methods have been developed to explore the functional pathways involved in tumorigenesis. Some of them focus on differential gene expression between healthy and pathologic tissues (89,90), on protein-protein interaction

network analysis(61,62)) or on molecular dynamics simulations(54). Other methods approach the disease through discretization of pathological components that result in tumo(91). All of these approaches are very powerful when the variables related to the disease, although complex, are well known and studied. A multi-factorial disease can be approached by means of mathematical theory, building a theoretical model where cell components are connected with each other. In biology, several problems were dealt with network theory (47,48). A network is a group of objects strongly inter-connected with each other (e.g. proteins and enzymes of a pathway or animals belonging to interacting populations). Their construction and subsequent simulation is made via mathematical analysis of the connections between nodes found in the system and their time-dependent behaviour(48)). A biological network is generally composed of proteins, nucleic acids and cofactors connected by biological reactions such as protein complex formation or enzyme activity regulation (48). Von Hippel-Lindau syndrome (VHL) (94) is a good study case to test the network theory applied to cancer due to the similar medical history and pathological phenotype that patients share. While hereditary cancers represent only a small part of all human tumors, their investigation represents a challenge to understand the pathway leading to tumor formation. In 2010, Heiner et al. first approached VHL using the so-called Petri Net (PN) simulation networks (95). Their work, inspired by a previous theoretical model of cellular oxygen-related pathways ((96,97), was a preliminary investigation of the core oxygen sensing system and its connection with VHL onset. Heiner and coworkers proposed three different functional modules responsible for hypoxia network control and for HIF-1α degradation (51). In other words, they theorized that hereditary forms of cancer, such as different manifestations of VHL, are the result of different and concomitantly compromised metabolic pathways.

3.3.1 Von Hippel-Lindau Disease

36

Von Hippel-Lindau protein (pVHL) is the product of the von Hippel-Lindau gene, located in the short arm of 3rd chromosome, and constantly transcribed in both fetal and adult tissues (4). Mutations of pVHL are related to a pathological outcome termed VHL syndrome, an inherited form of cancer (3). VHL syndrome is characterized by cysts and tumors growing in specific parts of the organism (3,12). It is considered a severe autosomal dominant genetic condition with inheritance of one person in over 35,000 (15). The tumor injuries, which can be either benign or malign, are usually located in the retina, adrenal glands, epididymis, central nervous system, kidneys and pancreas(56). As a genetic disorder, VHL syndrome follows Knudson's two hit principle. A copy of the gene is mutated in the germ line, but the other gene copy still produces a functional protein. Complete protein inactivation appears during life due to somatic inactivation of the remaining functional copy (17). On the contrary, mutations occurring during early fetal formation result in unsuccessful development(18). The pVHL gene has 11,213 base pairs including three exons (15) and the final transcript is a protein commonly present in two isoforms: pVHL30 and pVHL19, of 213 and 160 residues respectively. Neither isoform contains a known enzymatic domain, but rather appears to serve as a multipurpose adapter protein engaging in multiple protein-protein interactions (5). pVHL structure is organized in an α- and β-domain and its stability was demonstrated to be ensured by direct interaction with other proteins such as Elongins B and C(98). Both Elongin B and C are also required for the best characterized function of pVHL, the ubiquitination dependent degradation of Hypoxia Inducible Factor (HIF) via the proteasome (13). However, pVHL is considered a multipurpose protein due to its high number of known interactors. At the time of writing, the IntAct database(99)) presents more than 200 different interaction partners, with some of them competing for the same Elongin binding site. Indeed, pVHL was found in different cellular compartments and seems to be involved in many different cellular processes such as apoptosis, cell proliferation, survival and motility (100). Considering the huge number of interactors and

multiple cellular localizations, many different functions have been described or hypothesized, such as regulation of cytoplasmic microtubules during mitosis (60) and endothelial extracellular matrix deposition (101). On the other hand, considering the huge number of players involved in VHL syndrome and the lack of reliable kinetic data, a PN based approach may be a preferable option for an entire VHL pathway simulation.

3.3.2 Petri Net for Interaction Pathways

Since their invention, by Carl Adam Petri in the early sixties, PNs were mostly used to describe technical systems, but later the utility in describing biological and biochemical functions has also been demonstrated (53). PNs were successfully used in many studies to describe biological networks ((49) such as the regulation and etiopathology in human Duchenne Muscular Dystrophy (51) and the hypoxia response network (95). PNs are qualitative mathematical models that can graphically represent many object types, not only metabolites but also different protein states and are useful to simulate networks where not only metabolites are involved. Indeed, PNs can be a powerful tool to study all concurrent interactions in a specific pathway, even if the proteins or kinetics are not well-known. Due to the large number of different pVHL functions involved in VHL disease progression, we decided to extend the PN based analysis of (95) increasing the number of considered protein-protein interactions. We generated a novel manually curated PN model of the entire VHL regulation system collecting data from the literature and including the signaling pathways and glucidic metabolism. In order to build a realistic network, literature from both biochemical experiments and *in silico* predictions were used as source. It was decided to build a PN with only confirmed pVHL interactions whose function was also known. The resulting PN was validated using an analysis of specific properties as suggested by previous studies using the same method ((53). After validating the PN structure, *in silico* knock outs

of specific proteins were done in order to observe the different network behaviors and the resulting biological effect.

3.4 Methods

The network was designed in the Snoopy PN framework (version 2, revision 1.13) (102), respecting the mathematical PN formalism as described in (95,103). PN were demonstrated to be useful in describing discrete and concurrent processes in a simple graphical representation (49) and have been used to describe biomedical processes due to their capacity of representing sequential steps in a process. PN modeling methods are actively used to describe, simulate, analyze, and predict the behavior of biological systems. The Snoopy PN framework provides an extensible multi-platform framework to design, animate, and simulate Petri nets (102). We chose Snoopy to facilitate future extensions of the VHL pathway presented here. Among different available PN types a standard PN was chosen to limit the number of variables. Both Charlie and PInA analyzers were used for PN analysis and validation (104). Further, *in silico* knock out experiments were used to test the biological reliability of the model. Structural model validation was made by analysis of the T-invariants to demonstrate whether the system was covered by T-invariants and to confirm the biological meaning of each invariant. The use of T- and P-invariants is given by their own properties: they are a set (of transitions or places, respectively) that allows the reproduction of the same state after n transformations. A P-invariant represents a set of places where the number of tokens is constant and independent on the firing rate. A T-invariant instead represents a set of transitions that cyclically comes back to show the same initial set. Biologically a P invariant can represent the process of regulating a protein, whereas T invariants can represent cyclical biochemical transformations such as metabolic reactions. To this end, the computed invariants were grouped in Maximal

Common Transition Sets (MCTS) and Clusters, the former based on occurrence of specific sets of transition inside the various T-invariants, and the latter based on similarities between T-invariants. Different numbers of clusters will be defined depending on the resulting square matrix. Where MCTS create disjunctive nets, Clusters merge together similar T-invariants. Behavioral validation was made by selectively deleting tokens inside the model, imitating possible biological disruptions such as disease-causing mutations. The resulting network behavior was compared to what is reported in the literature. Total runtime for invariants computation were less than ten seconds on a mainstream Linux x86 workstation. Literature sources used to build the model are reported in Table S1 (see Appendix). The Snoopy framework for PN construction, Charlie and PInA tools for analysis are available at the website (URL: http://www-dssz.informatik.tu-cottbus.de /DSSZ/Software). Finally, the model was used to simulate the network behavior through visual inspection of both token movement and accumulation in specific parts of the network. For a visual explanation of token movement in a PN refer to Video S1 (available only for the online version).

Model Availability

The resulting VHL disease PN model is available in File S1 (see Appendix).

3.5 Results

3.5.1 Notations and Assumptions

The PN built here focuses on pVHL interactions that were already proven by biochemical experiments and reported in the literature. We chose to model a realistic VHL disease pathways based on confirmed literature data, including all known VHL functions, VHL related signal pathway and glucidic metabolism. All

bibliographic sources used to design the model are presented in Table S1 (see Appendix). The final PN is composed of 323 places and 238 transitions, connected by 801 arcs. Tables S1 and S2 (see Appendix) show all places and transitions and the related biological correspondence. Places are mainly proteins and enzymes, while some represent DNA or small molecular substrates such as glucose and cofactors (e.g. ATP). Notation for both pre- and post-places and their biological meaning are explained in Table S1 (see Appendix). In a few cases, places are used to represent a whole group of changes generated by DNA transcription, (e.g. p_32 and p_33 or Et_eff1 and Et_eff2). Transitions instead symbolize complex formation between two proteins or post-translational modifications. Output transitions stand for degradation or movement to other parts of the cell or organism to complete their functions (e.g. degrad_1 and degrad_2) whereas input transitions show the

generation of a substrate or protein. In order to simplify the design of such a large network, we decided to use macro nodes to group reactions representing complex molecular pathways such as signaling pathways or secondary signal cascades. The whole process is merged into a single node with a given name to allow visual inspection only in case of need. From the top level all transitions can still be found in a hierarchical lower layout level. Logic nodes were used for places participating in many reactions throughout the network such as ATP and ADP (7 logical copies each) or NAD and NADH (4 logical copies each). A total nesting depth of two was chosen to model macro nodes. Special arcs were not used while we chose to model the permanent presence of some objects using double arcs (e.g. for elob, eloc and places standing for enzymatic activity). In case of proteins, which are actively degraded, it was preferred to create an input transition simulating constant

production (or synthesis) and an output for consumption. This is the case for pkcz2, Jade1, pVHL and HIF-1α. As can be seen from figures 7 to 10, which
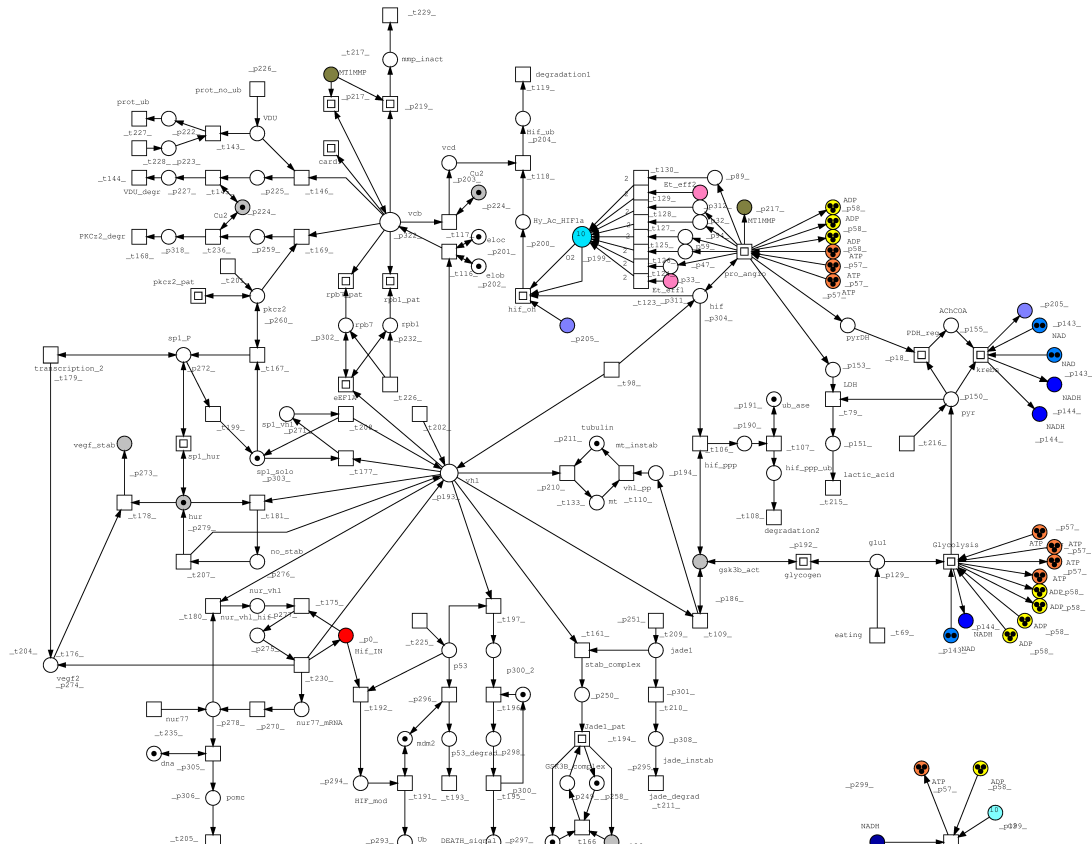
*Figure 7: **Lower hierarchical PN levels.***
Pathways from the top level are grouped in macro-nodes (functional subordinated layer), in particular glucidic metabolism and various VHL functions.

represent the entire model, two major nodes can be immediately identified: pVHL and vcb, the complex made by pVHL and the two elongins. Another relevant part is the glucidic metabolism, modeled due to its hypoxia-induced regulation. It is represented in detail in Figure 8.
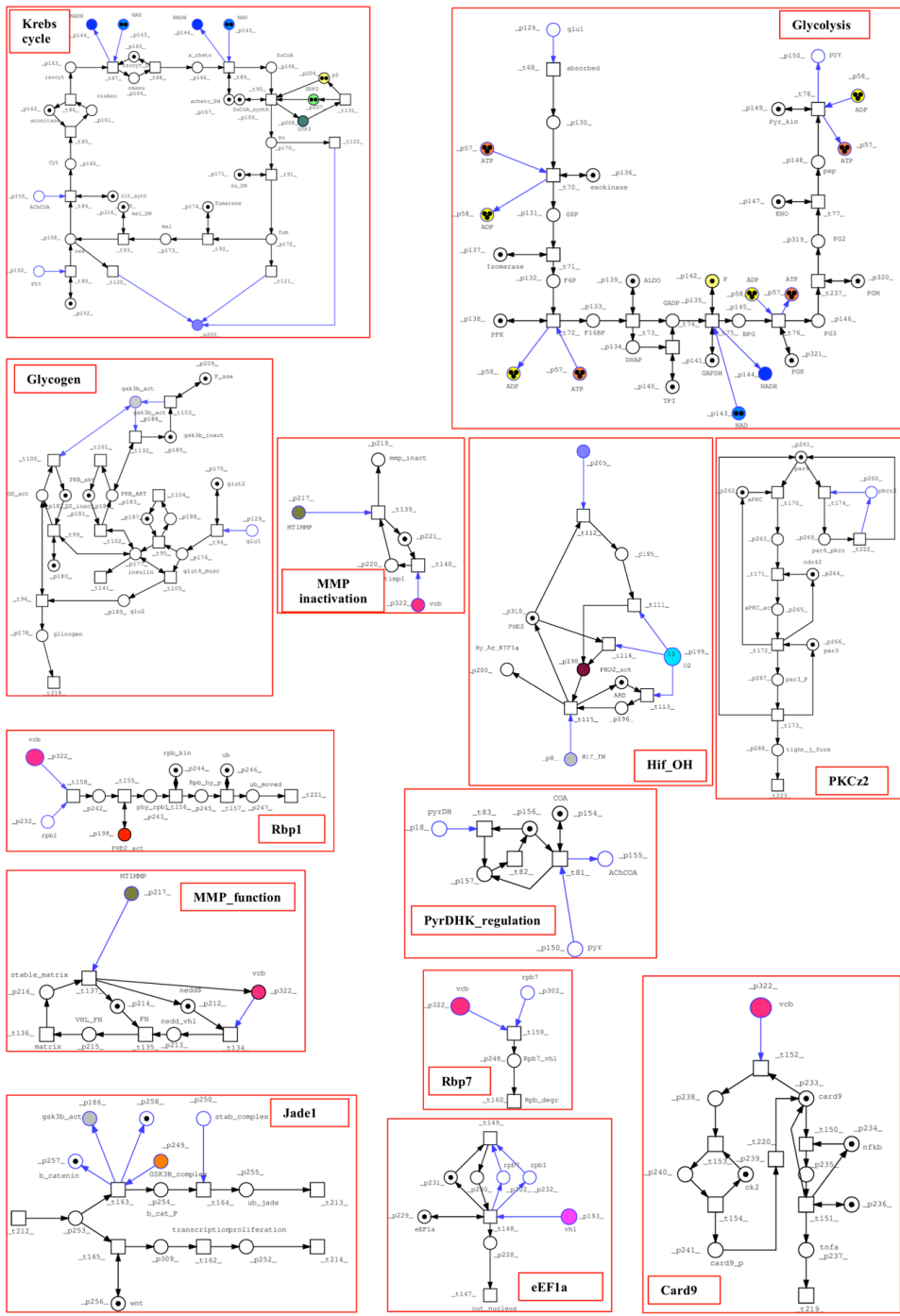
Figure 8: Lower hierarchical PN levels.
Pathways from the top level are grouped in macro-nodes (functional subordinated layer), in particular glucidic metabolism and various VHL functions.

44
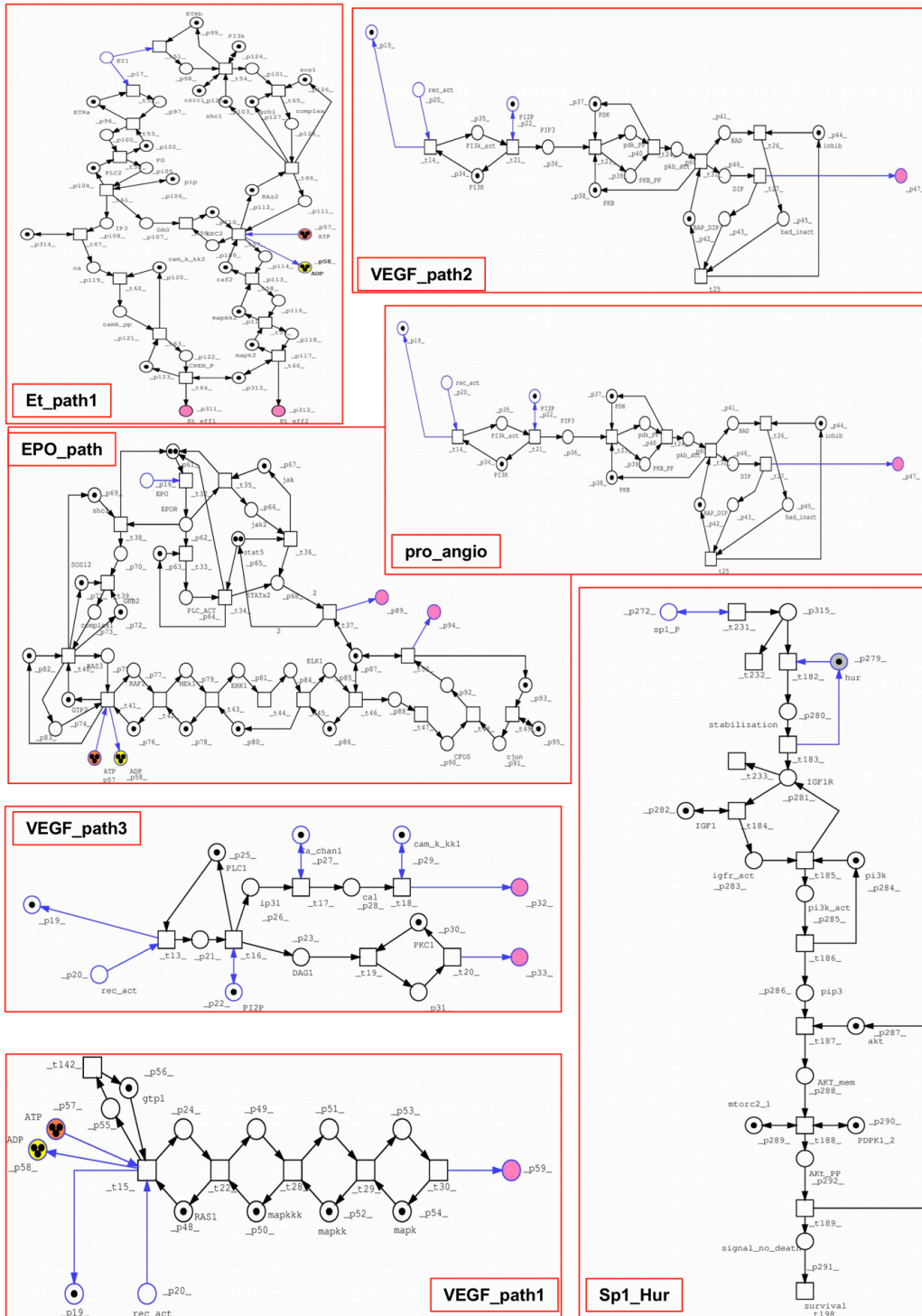
*Figure 9: **Lower hierarchical PN levels, in particular HIF-1α regulation and HIF-1α-dependent pro-angiogenic signaling.***
VEGF and EPO pathways are at a lower hierarchical level than the pro_angio macro-node.

3.5.2 HIF-1α Transcription Activity

The HIF-1α transcription factor stimulates proliferation of endothelial cells to create new blood vessels during localized or broad hypoxia. In human, it is present as three different paralogs: HIF-1α, HIF-2α and HIF-3α. The sequence is quite conserved between the former two, whereas the latter is slightly shorter and seems to have completely different functions compared to the other two(103,104). Both HIF-1α and -2α stimulate DNA transcription but the exact products of this activity are still poorly understood. In our model, only HIF-1α *in vivo* activity was considered. It cannot be excluded that other biological effects depend on the second paralog. Indeed, both have a pro-angiogenetic function and are degraded by pVHL via proline-directed hydroxylation. HIF is a heterodimer of HIF-1α and HIF-1β, the latter being also termed Aryl hydrocarbon Receptor Nuclear Translocator (ARNT). We started from the transcription activity of HIF due to its regulation is the most studied pVHL function. Our model, as expected from literature data, shows that HIF-1α enters the nucleus when not degraded by pVHL. It subsequently binds HIF-1β to form the HIF heterocomplex which interacts with DNA. Our model correctly simulates the increased affinity of HIF towards DNA. Transcription is enhanced by some co-factors binding both subunits of HIF and other proteins such as p300, Creb and cjun. This takes place in a specific DNA promoter sequence termed Hypoxia Response Element (HRE). Furthermore, during transcription some pro-angiogenic factors are produced: Vascular Endothelial Growth Factor (VEGF), Endothelin (ET) and Erythropoietin (EPO). All described pathways are in agreement with previous observations reported in (105).

3.5.3 Metabolic Processes

HIF-1α transcription activity includes some proteins which are dependent on oxygen but involved in other pathways (e.g. oxidative metabolism) or completely

independent (e.g. metallo-proteinase MT1MMP). Further, HIF-1α stimulates production of proteins involved in the glucidic pathway. The final product of the metabolism is adenosine triphosphate (ATP), a molecular form of energy, composed by adenosine, an adenine ring connected to a ribose sugar, and three phosphate moieties. When a phosphate moiety is hydrolyzed it releases energy, used by cells for enzymatic reactions. The glucidic metabolism is composed of glycolysis, Krebs cycle, glycogen formation and respiratory chain with ATP synthesis. Glucose is absorbed in cells by enzymatic glucose transporters (GLUT), which carry the molecule to the location inside the cell where the metabolism takes place ((106). There are many isoforms of these transporters: GLUT1 is present in all cells and in particular in erythrocytic membranes, neurons and glia (107). GLUT2, located in both liver and pancreatic beta cells, is characterized by low affinity for glucose, hence it requires a higher glucose concentration to be activated(108). Right after eating, glucose concentration increases, thereby quickly activating them. GLUT2 stimulates production of insulin, a hormone regulating the plasmatic glucose concentration. Glucose plasmatic concentration can also increase due to an opposite pathway, originating from liver glycogen being decomposed into glucose and reaching systemic circulation. GLUT3 is mostly present in neurons, whereas GLUT4 is the insulin activated transporter located in myocytes, adipocytes and cardiomyocytes (106,109). In our model, we chose to exclude GLUT3 due to its specific role in neuronal cells. Glycolysis occurs in the cytoplasm and during this process each glucose molecule is phosphorylated, consuming two molecules of ATP, then divided into two smaller molecules. Further modifications of these two molecules result in new ATP production. The molecule obtained at the end of glycolysis is pyruvate, which can be again modified through three different pathways. It can be decarboxylated and linked to Co-enzyme A to form acetyl-Co-enzyme A. It can then be carboxylated to obtain oxalacetate, or transformed through lactate dehydrogenase into lactic acid. Pyruvate can also be generated by other metabolic pathways, like protein or fatty acid disruption and amino-acid

modifications. Acetyl-CoA and oxalacetate are the molecules used in the following glucidic metabolism process, the Krebs cycle, taking place in the mitochondrial matrix. The Krebs cycle starts with acetyl-CoA and oxalacetate merging to create citric acid, which continues undergoing modifications until oxalacetate is formed again. During the process some co-enzymes are modified. Decarboxylation of pyruvate to form acetyl-CoA already transforms a NAD^+ (Nicotinamide Adenine Dinucleotide) in NADH (reduced form), afterwards obtaining one more of ATP, GTP, $FADH_2$ (Flavin Adenine Dinucleotide) and three more NADH per pyruvate molecule entering the Krebs cycle. The redox co-enzymes are considered electron transporters. During metabolic reactions they reduce themselves and get electrons (and protons) to oxidize the substrate of the enzymatic reaction. Electrons taken during the glucose metabolism are then used in the respiratory chain taking place in the internal mitochondrial membrane. The respiratory chain consists in transporting electrons through enzymes called cytochromes and others co-enzymes, characterized by the capability to receive and donate electrons. NADH ($FADH_2$) is oxidized again by cytochromes going back to the form of NAD (or FAD). Electrons gained through oxidation are used to reduce half a molecule of oxygen into water, releasing more energy. The $FADH_2$ and NADH redox chain establishes a chemical potential causing the push of protons outside the internal membrane towards the inter-membrane space, which stays between the mitochondrial inner and outer membrane. This also causes a higher concentration of protons outside the inner membrane. The resulting gradient causes the tendency of protons to enter the cell. The final step is ATP-synthetase, formed by a channel that allows protons to enter, pushed by the gradient, allowing the enzyme to change conformation and make its reaction. This kinetic energy is converted into ATP. In our model, the glycolytic and Krebs cycles were described in detail, represented at the hierarchical second level by the coarse transition Glycolysis. The respiratory chain was instead merged into a single node (t_97). We chose to represent creation and consumption of ATP in order to show the effects of lower and higher oxygen

48

concentration on the network. On the other hand, oxygen consumption for ATP synthesis during the respiratory chain creates a flow of oxygen in the model. Oxygen is not the only connection between glucidic metabolism and hypoxia. Indeed, HIF-1α transcription activity enhances the transcription of many GLUT isoforms (such as 1, 3 and 9) and the pyruvate dehydrogenase kinase, which determines the pyruvate dehydrogenase (PyrDH) inactivation and consequent Acetyl-CoA formation from pyruvate. Finally, Lactate dehydrogenase is also produced, to ensure an alternative compound, creating energy needed for cell survival (106,110).

3.5.4 pVHL-dependent Processes

Some interactors can bind pVHL in regions interacting with Elongin C. These are HuR, Nur77, p53 and Jade1. Nur77 has a complex function and its role in pVHL tumor suppressor activity is still not entirely clear. Nur77 can bind pVHL, inhibiting Elongin binding while allowing HIF-1α binding. Its transcription is stimulated by HIF-1α itself, and pVHL-HIF-1α-Nur77 complex formation stabilizes the transcription activity of HIF-1α by inhibiting the pVHL-dependent degradatio((111)). Another Nur77 function is the stimulation of proopiomelanocortin (POMC) transcription, which is a precursor for adrenocorticotropic hormone (ACTH) formation. This hormone has an important stress response function, stimulating cortisol production and other neurotransmitters from the adrenal glands, to enhance the organism reaction to danger and stress *stimuli* e.g. increase of gluconeogenesis and muscle mass. An excess of this hormone can cause desensitization of its receptors for feedback down-regulation and thus muscular weakness, tiredness, hyperglycemia and osteoporosis (112). p53 can bind to pVHL avoiding the degradation of this tumor suppressor. Instead, it stimulates the apoptotic signal cascade via the p300 co-activator, which stimulates production of proteins enhancing the cell-programmed

death. If p53 cannot bind pVHL, two more mechanisms are described in the model. One is its modification and degradation by Mdm2 and the other is the pVHL-independent degradation of HIF-1α. Interaction with Mdm2 is needed in both cases(113,114). Jade1 is a short-lived protein whose main function is to stimulate the phosphorylation-dependent degradation of β-catenin. This is a subunit of the cadherin protein complex acting as an intracellular signal transducer in the Wnt signaling pathway. It seems that β-catenin is able to stop cell division via a contact-dependent inhibition signal, whereas in Wnt signaling it is also involved in proliferative transcription. When Wnt is not present, β-catenin can be phosphorylated by Glycogen Synthetase Kinase, type 3β (GSK3β) in complex with APC (Adenomatous Polyposis Coli) and Axin. β-catenin can interact with Jade1 and be only successfully degraded after this interactio(115). Related functions are represented in the macro node Jade1_pat. GSK3β seems to be a protein involved in many different pathways. GSK3β is involved in Glycogen Synthetase deactivation and can even phosphorylate pVHL and HIF-1α. In the case of HIF-1α, it generates a pVHL-independent degradation pathway, where phosphorylation allows ubiquitination, whereas in the case of pVHL, it inhibits pVHL stabilization of microtubules (116).

3.5.5 Structural Model Analysis

Based also on previous observations of Heiner et al., (117), in 2008 Grunwald et al., demonstrated that PN can be used to describe large and complex metabolic pathways (51). They postulated the following set of minimal rules that a PN should satisfy to be considered biologically reliable: (i) the network should be entirely connected, (ii) the network should be covered by T-invariants, and (iii) each T-invariant and P-invariant should have a biological meaning. The model described here was tested with respect to what previously done by Grunwald and co-workers (51) and resulted to be covered by T-invariants, connected, homogeneous

and each place has a pre-transition and a post-transition. Transitions without pre- or post-places were used to simulate the system interface to the surroundings. The network is alive, in other words, it continues to work forever, with all transitions contributing to the net behavior forever, and no dead transitions. The MCTS and Cluster analysis were used due to the large number of T- and P-invariants included in the model. Both methods are used in PN theory to reduce the complexity connected with such a large network and to reduce the errors connected with manual investigation. From the 238 transitions present at the beginning in the model, 393 T-invariants were computed without considering 10 trivial invariants. The latter consist in a pair of transitions that usually represent a forward and backward reaction, such as the active and inactive state of a protein. Trivial invariants could be erased to reduce the dimension of the network without disturbing the overall system when the interest is focused on the steady state behavior (95). T-invariants were grouped into 44 Clusters using the Tanimoto coefficient with similarity threshold of 65%, as described in (51). Only 11 of these 44 comprised more than one T-invariant. The three biggest Clusters are C9, composed of 144 T-invariants, C8 of 72 and C11 of 64 T-invariants. Separation into clusters allows easier analysis of networks pathways represented by each T-invariant, since they are grouped by similarity, specifically the common transitions by which they are composed. T-invariants named in the text are shown in Table S3 (see Appendix), while T-invariants grouped in C8, C9, C10, C11 are explained in Table S4 (see Appendix) and described as follows.

3.5.5.1 Cluster C8

Cluster C8 groups all transitions included in HIF-1α pathways, including transcription, signaling cascades, degradation via pVHL, p53 and GSK3β, and eventually the Krebs cycle. For the EPO signaling pathway, two transitions (t_35 and t_36) are not included which cause Jak activation and consequent Stat5 activation to stimulate DNA transcription. Matrix stability regulation is also part of

the cluster due to the destabilization induced by HIF-1α transcription of metallo-proteinase (MMP), transitions from t_134 to t_140. The largest T-invariant in C8 is Inv_280 (93 transitions) while the smallest is Inv_377 (81 transitions). The differences between T-invariants show the possibility of alternative pathways inside the model. For example, the VEGF dependent signal cascade can proceed in three different ways: t_13, t_14 and t_15, which lead to the pathways being merged in the coarse nodes Vegf_path3, Vegf_path2 and Vegf_path1, respectively. The occurrence rate in C8 is 24 transitions for each path. The Endothelin, VEGF and Erythropoietin pathways are not in conflict and occurring together. Disaggregation of the matrix via MMPs is present in 18 T-invariants, whereas inhibition of these proteins, i.e. matrix stabilization, is present in the remaining 54 transitions. Regarding the Krebs cycle, 47 T-invariants have t_91, of which only 24 reach t_92 and t_93, representing the last three steps of the cycle: succinate to fumarate, fumarate to malate, and malate to oxalacetate. All the malate being produced is used to regenerate oxalacetate. Degradation of HIF-1α occurs in any T-invariant of the cluster. The pVHL-dependent degradation of HIF-1α is always present (transitions t_116 to t_119). In 19 T-invariants degradation takes place via p53 (t_191 to t_193) or, alternatively, via phosphorylation by GSK3β in another 17 T-invariants. Two of the three pathways can be present in the same T-invariant, as in Inv_227, where degradation via pVHL and degradation via p53 are both present. This was considered as the HIF-1α dependence on the lack of degradation by these proteins. All three degradation pathways never appear in the same T-invariant. The p53 and GSK3β paths are never present together but each of them is accompanied by pVHL-dependent proteasomal degradation. Inv_377 lacks the EPO signaling pathway but is the only one in this cluster to have t_34, t_33 and t_37. These invariants have all input and output transitions. For example, t_202 the second input for pVHL, is present in only 18 invariants. Other inputs are t_98, always present, leading to formation of HIF-1α and pVHL, t_192, producing p53 and t_216, representing other pyruvate generating metabolic pathways. The latter is also

present in each invariant allowing formation of the pyruvate needed for Krebs cycle progression.

3.5.5.2 Cluster C9

Cluster C9 is the largest cluster in our model and includes 144 T-invariants. It is characterized by complete EPO pathway abrogation which goes through formation of the Shc-Grb-Sos complex and the consequent mapk-dependent phosphorylation cascade. Transition t_127, representing EPO effects on oxygen production, is absent. In its place, t_35 and t_36 are considered, which are present in 72 T-invariants. In cluster C9, the largest T-invariants are Inv_278 and Inv_279 (74 transitions) while the shortest ones are Inv_101, Inv_105, Inv_144 and Inv_148 with 65 transitions each.

3.5.5.3 Cluster C10

Cluster C10, composed of 52 T-invariants, is characterized by the presence of glycolysis between many transitions grouped in the cluster. This is also the cluster containing the most populated T-invariant of all computed 393 non-trivial T-invariants. This is Inv_245, including 101 transitions and covering almost half of the whole model. Cluster C10 also includes Inv_125, the shortest invariant of this model, composed by 85 transitions due to lack of the Krebs cycle. Another difference with the other three major clusters is that here both EPO paths are present, specifically, the Jak pathway belongs to 4 T-invariants and Shc-Grb-Sos is observed throughout the cluster. Vegf_path1 seems to be more common in this cluster, being present in 36 T-invariants, whereas the other two are present 12 times each. This time they are present even in the same invariant, as for Inv_60, Inv_129, Inv_172 and Inv_215, with both t_13 and t_15, and Inv_142, Inv_185 and Inv_228 with t_14 and t_15 and all subsequent signaling appearing at the same time. Despite glycolysis being present in all cluster invariants, the Krebs cycle appears only in 11 cases. p53-dependent degradation of HIF-1α occurs in 11 cases

while the phosphorylation-dependent one appears in 13. An input transition has been added with respect to the other major clusters so far analyzed (i.e. t_69_eating) without which glycolysis could never take place.

3.5.5.4 Cluster C11

Cluster C11 is composed of 64 T-invariants. Only part of the EPO pathway is described here, with the major difference that the Krebs cycle is completely abrogated while Prolyl Hydroxylase type 2 (PHD2) regulation by oxalacetate is included. HIF-1α interaction with Nur77 and transcription of VEGF by Sp1 are also present. t_80 (transformation of pyruvate in oxalacetate) is not present in the first 42 cluster T-invariants. Nur77 interaction with HIF-1α is present only in 8 T-invariants, specifically Inv_87 to Inv_94. Sp1 transcription activity is appearing in twice the amount, including the same 8 invariants just mentioned. VEGF transcription via Sp1 activity is aPKCζ2 phosphorylation dependent, which does however not appear in the cluster. When VEGF is synthesized, it is subsequently stabilized by Hur, followed by t_178 and Hur is recreated to allow other functions. Indeed, it is one of the few places without input transition but with a token that goes forward and backward again. Compared to the other clusters, C11 also shows one less transition in the coarse PHD regulation node, specifically t_81, which shows the transformation of pyruvate by pyrDH into acetyl-Coenzyme A, needed for the Krebs cycle. The four clusters C-8 to C12 are very similar to each other, as can be seen from the distance tree in Figure 10. They all contain the HIF-1α transcription activity and signaling pathways caused by EPO, VEGF and the HIF-1α degradation options. They include the effects of other transcription activity products, like metallo-proteinase and pyruvate dehydrogenase kinase, which regulate the activation state of PyrDH. All include part of the glucidic metabolism but not Glycogen formation itself. Other five clusters from C12 to C16 have a smaller number of T-invariants and fewer transitions present in each invariant. They do not include transcription activity but are only formed by the VEGF and

glycolytic pathways. The information contents of these clusters turned out to be uninformative and their analysis was not included. The same applies to clusters composed by 1–3 T-invariants. Finally, some transitions are not present in the clusters and not listed in the T-invariants because trivial invariants were excluded from cluster analysis. These transitions are shown in Table 1 with their respective biological meaning.
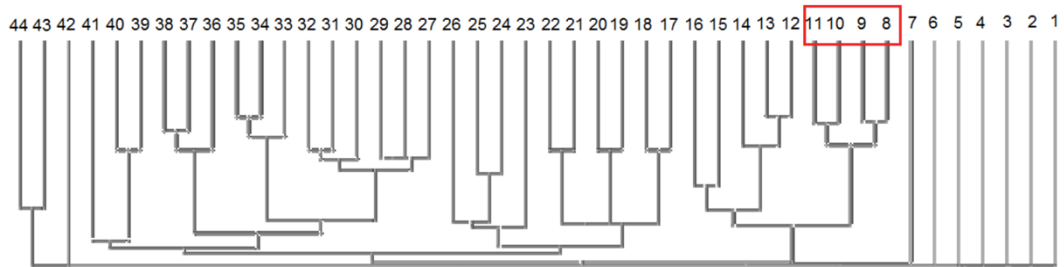
*Figure 10: **PinA Distance Matrix clustering, using Tanimoto coefficient and 65% threshold.** The numbers indicates clusters. In C8, C9, C10, C11 are highlighted a red square.*

Trivial T-Invariants	ID transitions	Biological Meaning
TInv_1	t_99, t_100	Glycongen Synthase regulation
TInv_2	t_101, t_102	Pkb regulation
TInv_3	t_103, t_132	GSK3β active-inactive state
TInv_4	t_174, t_222	Par6 inactivation via aPKCζ2
TInv_5	t_177, t_208	VHL binding to Sp1
TInv_6	t_167, t_199	Sp1 phosphorylation and dephosphorylation
TInv_7	t_0, t_2	Hif transport in and out of nucleus
TInv_8	t_0, t_234	Hif inhibition via FIH
TInv_9	t_181, t_207	Hur inhibition via VHL
TInv_10	t_231, t_232	IGFR mRNA production and destruction

Table 1: List of Trivial T-invariants excluded from calculation with their associated biological meaning.

3.5.6 MCTS Analysis

Another way to group invariants is by the amount of single transitions present in them. Maximal common transition set (MCTS) analysis provides a PN decomposition into non-overlapping subnets, sharing parts of the same T-invariants (53). In a biochemical network, MCTS could be interpreted as enzyme subsets operating together under steady state conditions, computed based on the support of a T-invariant. MCTS computation does not consider stoichiometric relations, describing exclusively sets of reactions present in a maximal number of T-invariants resulting shared by different signaling pathways (118). A total of 40 non-trivial MCTS were identified, with results and related biological means shown in Table 2 (see Appendix). Some transitions do not belong to a non-trivial MCTS, because their occurrence has no similarity with other transitions and they create separate MCTS (specifically: t_69, t_82, t_91, t_94, t_98, t_114, t_116, t_120, t_121, t_122, t_179, t_202, t_209, t_212, t_216, t_225 and t_229). MCTS define transitions that always take place together, but are not necessarily connected, thus representing disjunct building blocks constituting the network. Considering both analyses, a table was automatically built in PInA (53) showing a correlation between clusters and MCTS. Transitions (t) or MCTS (M) are compared to evaluate how many T-invariants clusters cover the selected M or t (if the transition is not already part of the MCTS, as listed above). The more covered a transition or set is, the more central it could be considered for the network behavior. Recently, a network coarsening method based on abstract dependent transition sets (ADT) was presented ((119). It is formulated without the requirement of pre-computation of the T-invariants and is a tool commonly used for the decomposition of large biochemical networks into smaller subnets. Due to the manually designed nature of our model, we preferred to maintain a logic hierarchy based on metabolic pathways in order to maintain the network centered on pVHL and its interaction. The MCTS calculation results shows that the most covered set by cluster T-

invariants is M20 with 358 T-invariants covering all transitions in the set, indicating that this MCTS corresponds to more T-invariants than the others. All transition sets are an important link to the others, as tokens pass through these transitions more often. A transition not present in any set but most covered by T-invariants is t_98, which is also the most frequently occurring transition, see Figure 11. The 10 most occurring transitions are listed in the Table 3.
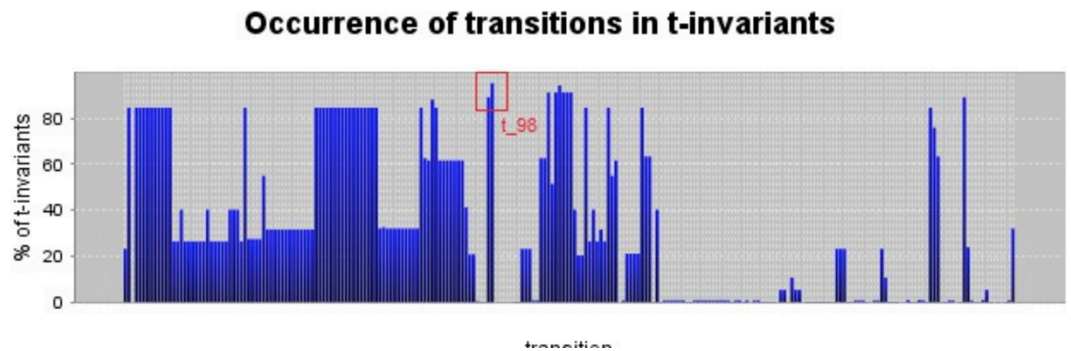
*Figure 11: **Transitions occurrence T-invariants.***
Transitions are ordered by name and t_98 is highlighted in red.

MCTS	ID Transitions
MCTS 1 (M1)	t_0, t_190, t_191, t_192;
MCTS 2 (M2)	t_1, t_3, t_4, t_5, t_6, t_7, t_8, t_9, t_10, t_11, t_12, t_32, t_51, t_52, t_53, t_54, t_55, t_56, t_57, t_58, t_59, t_60, t_61, t_62, t_63, t_64, t_65, t_66, t_67, t_79, t_83, t_123, t_129, t_138, t_215;
MCTS 3 (M3)	t_2, t_99, t_100, t_101, t_102, t_103, t_132, t_167, t_174, t_181, t_199,_177, t_207, t_208, t_222, t_232, t_234;
MCTS 4 (M4)	t_13, t_16, t_17, t_18, t_19, t_20, t_124, t_128;
MCTS 5 (M5)	t_14, t_21, t_23, t_24, t_25, t_26, t_27, t_31, t_126;
MCTS 6 (M6)	t_15, t_22, t_28, t_29, t_30, t_125, t_142;
MCTS 7 (M7)	t_33, t_34;
MCTS 8 (M8)	t_35, t_36;
MCTS 9 (M9)	t_37, t_130;
MCTS 10 (M10)	t_38, t_39, t_40, t_41, t_42, t_43, t_44, t_45, t_46, t_47, t_48, t_49, t_50, t_127;
MCTS 11 (M11)	t_68, t_70, t_71, t_72, t_73, t_74, t_75, t_76, t_77, t_78, t_237;
MCTS 12 (M12)	t_80, t_111, t_112;
MCTS 13 (M13)	t_81, t_84, t_85, t_86, t_87, t_88, t_89, t_90, t_131;
MCTS 14 (M14)	t_92, t_93;
MCTS 15 (M15)	t_95, t_104, t_141;
MCTS 16 (M16)	t_96, t_105, t_218;
MCTS 17 (M17)	t_97, t_224;
MCTS 18 (M18)	t_106, t_107, t_108;
MCTS 19 (M19)	t_109, t_110, t_133;
MCTS 20 (M20)	t_113, t_115, t_117, t_118, t_119;
MCTS 21 (M21)	t_134, t_135, t_136, t_137;
MCTS 22 (M22)	t_139, t_140, t_217;
MCTS 23 (M23)	t_143, t_227, t_228;
MCTS 24 (M24)	t_144, t_145, t_146;
MCTS 25 (M25)	t_147, t_148, t_149;
MCTS 26 (M26)	t_150, t_151, t_219;
MCTS 27 (M27)	t_152, t_153, t_154, t_220;
MCTS 28 (M28)	t_155, t_156, t_157, t_158, t_159, t_160, t_221, t_226;
MCTS 29 (M29)	t_161, t_163, t_164, t_166, t_213;
MCTS 30 (M30)	t_162, t_165, t_214;
MCTS 31 (M31)	t_168, t_169, t_201, t_236;
MCTS 32 (M32)	t_170, t_171, t_172, t_173, t_223;
MCTS 33 (M33)	t_175, t_176, t_180, t_230;
MCTS 34 (M34)	t_178, t_203;
MCTS 35 (M35)	t_182, t_183, t_231, t_233;
MCTS 36 (M36)	t_184, t_185, t_186, t_187, t_188, t_189, t_198;
MCTS 37 (M37)	t_193, t_194;
MCTS 38 (M38)	t_195, t_196, t_197, t_200;
MCTS 39 (M39)	t_204, t_205, t_206, t_235;
MCTS 40 (M40)	t_210, t_211;

Table 2: *List of MCTS and transitions from PInA.*

Rank	Transitions	Biological meaning	Occurrence %
1	t_98	Input transition for Hif and VHL	95.165
2	t_116	Interaction of VHL with Elongin B and C	94.148
3	t_113	Activation by oxygen of ARD	94.094
4	t_115	Acetylation and hydroxilation of Hif	94.094
5	t_117	Interaction of complex Vcb with Cu2	94.094
6	t_118	Interaction of complex Vcb with modified Hif	94.094
7	t_119	Degradation VHL dependent of Hif	94.094
8	t_97	ATP formation	89.059
9	t_224	Water Output transition	89.059
10	t_82	Pyruvate Dehydrogenase inactivation	88.041

Table 3: Ranking of the 10 most occurring transitions with biological meaning and percentage of occurrence.

3.5.7 P-invariant Analysis

Although the network is not covered by P-invariants, it has 130 P-invariants. 47 of these are trivial P-invariants, comprising a single place, connected with double arcs to imitate an activator arc function. Another object represented with double arcs is the enzymatic activity catalysing a reaction and immediately going back to the steady state. P-invariants show places or sets of places where token numbers always remain equal and do not move outside the subnetwork induced by the P-invariant in the initial marking. In other words, they do not grow nor diminish. The remaining P-invariants are mostly located in signal transduction pathways, such as situations in which a protein is sequestered from its function and then goes back after a second reactivation mechanism. This scenario is present in p_41, p_42 and p_45 located in invariant P_58. It is important to notice that ATP and ADP, as well as NAD and NADH, are modeled as P-invariants. P_90, P_91 and t_97 are able to transform ATP and ADP. More in general, all energy consuming transitions are considered to be backward transitions of invariants. Invariants not related to signal transduction are places located in the Hur system, where Hur is removed from its function by pVHL. This is a good approximation for sequential modifications that momentarily activate proteins. Afterwards, Hur can go back and stabilize VEGF to increase its transcription activity.

3.5.8 In Silico Knock Out Experiments

The previously described clustering and MCTS analysis for T-invariants allowed us to identify the most common transitions and to understand which transitions can be depleted in our knock out experiments in order to get the most important biological effect. The knock out experiments were performed erasing selected transitions or tokens and observing which transitions or MCTS become inactivated. Considering our results and the literature, we decided to knock out the following pathway elements: (i) pVHL, (ii) HIF1α alone and with Sp1, (iii) t_98, (iv) PHD2, (v) MCTS1, (vi) t_97 and (vii) GSK3β. In the following, we describe the effect of each knock out scenario on our model.

(i) pVHL knock out.

Degradation of HIF-1α is not completely depleted due to presence of both p53- and GSK3β-dependent alternative degradation pathways. All other processes usually inhibited by pVHL take place in an uncontrolled way, including creation of VEGF via Sp1 transcription activity and increased matrix regulation due to lack of fibronectin crosslinking. Hur resulted constantly activated and nur77 can stimulate synthesis of Proopiomelanocortin, precursor for the Adrenocorticotropic hormone. Card9 increases release of tumor necrosis factor, and NF-kB when not inhibited by pVHL. Instead, Jade1 is unable to survive long enough to inhibit β−catenin, generating a proliferation signal with Wnt. Lactic acid is also not produced due to LDH enzyme production being HIF-1α transcription activity dependent.

(ii) HIF-1α knock out.

VEGF is still created thanks to Sp1, thus oxygen is still generated even if in lower proportion. If HIF-1α and Sp1 are both knocked out at the same time, oxygen is quickly consumed and the metabolism is soon unable to proceed. Lactic acid is not

produced due to LDH enzyme production being HIF-1α transcription activity dependent. Glycolysis and glycogen are produced normally and the metabolism is not inhibited by PyrDH negative regulation and lactic acid formation. Since pVHL is present, other tumor suppressor activities are enabled, except for proteasomal degradation of HIF-1α due to the substrate being non-existent.

(iii) HIF-1α and pVHL double knock out.

This generates a situation where the metabolism is normal but oxygen regeneration is less productive, with only Sp1 acting for transcription. Due to absence of pVHL, all proliferation-stimulating processes are active, causing an unbalanced consumption of resources. Our model shows that this condition is compatible with cell growth and multiplication, but new blood vessel generation is consistently slower and glucidic metabolism appears principally based on the glycolysis reaction. Similar activity reduction applies to both tight junction and cellular external matrix (ECM) pathway regulation. It cannot be excluded that some observed effects could be mitigated by both HIF-2α and HIF-3α activity *in vivo.*

(iv) PHD2 knock out.

The protein is involved in pVHL mediated and oxygen dependent degradation of HIF-1α. Further, PHD2 is involved in hydroxylation of the RNA polymerase II subunit Rpb1 to allow its translocation to less chromatin-concentrated areas of the nucleus. When it is knocked out, HIF-1α degradation can continue via alternative pathways as seen in the pVHL knock out experiment and there is more RNA polymerase II activity, even if rpb7 can still be inactivated by pVHL.

(v) MCTS1 knock out.

MCTS1 groups some reactions involved in the HIF-1α p53-dependent degradation pathway (Table S2, see Appendix). To perform this knock out, we erased the

necessary token in mdm2, making the precondition insufficient to enable the MCTS transitions. p53 is not degraded and can continue its proapoptotic signal. On the other hand, a HIF-1α degradation mechanism is also knocked out resulting in an increased HIF-1α transcription activity.

(vi) t_97 knock out.

This is the ATPase transition, allowing the model to imitate oxygen consumption for ATP synthesis. If this transition is inactive, oxygen accumulates infinitely and ATP is not regenerated after few simulation steps. At the beginning, ATP is formed during the first step of glycolysis but afterwards it is consumed again. At some point, these reactions do not have any ATP available to allow the system to re-balance the consumed ATP. After few simulation steps, oxygen reaches a high level due to slower consumption in the PHD2 regulation process. Biologically, this means that the metabolism stops and the cell is not able to create energy to survive. There is no accumulation other than glucose in the model. A few oxygen creation processes are blocked as well due to absence of ATP, e.g. t_15, t_41 and t_57.

(vii) GSK3β knock out.

This enzyme is involved in negative glycogen synthetase (GS) regulation and is inactivated when phosphorylated. When GSK3β is knocked out, glycogen is continuously produced due to the enzyme remaining in an active state. In a real organism there are alternative forms of GSK3β which can inactivate GS, hence the effect will be less sharp. GSK3β is also involved in the degradation of HIF-1α, causing its phosphorylation and following ubiquitination. It is also involved in the degradation of β-catenin, where it is responsible for primary phosphorylation. If knocked out, even if Jade1 can be stabilized by pVHL, the effect will be similar to a knock out of Jade1, where β-catenin is free to continue proliferation stimulating transcription activity.

3.6 Discussion

We started from a core model of hypoxia response (95) and extended the original network with functional data derived from the literature in order to represent a complete description of the pVHL interaction pathway according to current knowledge. VHL syndrome is characterized by the formation of tumors and cysts affecting different organism districts and tissues. Indeed, pVHL is a tumor suppressor whose functions are connected to inhibition of proliferation and survival, growth and stability of extracellular matrix and microtubules, as well as cell polarity and migration. The IntAct database reports more than 200 suspected pVHL interactors and for most of them interaction and function details remain largely unknown. We chose to model the pVHL interactions in a credible cellular context with many protein activities occurring at the same time. The main idea was to create a novel manually curated PN description of the entire VHL disease pathway, including glucidic metabolism and signaling pathways. The model was designed as a standard PN and is composed of 238 transitions and 323 places, connected by 801 edges. A biologically realistic PN model needs to be covered by T-invariants, meaning each transition in the model has to be included in a T-invariant, and each invariant needs to have a biological meaning (51,117). We used the T-invariant analysis to validate the reliability of the model. We computed a total of 393 T-invariants, plus 10 trivial invariants, which were excluded from analysis. These were grouped into 44 Clusters and, through use of T-invariants, transitions were grouped into 40 MCTS. The model obtained is connected, covered by T-invariants with each invariant holding a biological meaning. MCTS analysis was used to identify the most frequent crucial transitions occurring in the model. This specific subset was further used to plan *in silico* knock out experiments and for the model validation and analysis of expected biological behavior. The model was then used to perform *in silico* knock out experiments inactivating specific transitions during qualitative network analysis. Our results showed that the model is able to

represent important transitions reflecting real biological outcomes, i.e. transitions involving species such as oxygen or ATP are correctly inactivated under certain circumstances as expected from the bibliographic data. Biological energy-related reactions (e.g. ATP production from ADP) were modeled as P-invariants. Although the network is intentionally not covered by P-invariants, P-invariant analysis was used to verify all modeled energy consuming transitions. Both the ATP and NADH balances appeared constant during the simulation, with irrelevant P-invariants located in the Hur system. This approximation was used to verify the Hur-dependent regulation of VEGF, with results in accordance with (120). The specific pVHL knock out suggests that this protein alone is not sufficient for complete HIF-1α inactivation. Indeed, other concurrent HIF-1α degradation pathways promote a sort of cell cycle regulation backup. On the contrary, simple deletion of pVHL turned out to be sufficient to increase all its other inhibitory functions, showing similar effects to pathological VHL symptoms. Indeed, ECM destabilization increases cell migration to other areas, promoting metastasis outbreak in case of tumor cells. Further, pVHL-dependent inhibition of tight junction formation by aPKCζII participates in an easier cellular detachment. The interactions of Nur77 could be considered a good example for pathological effects. It is a stimulator of Proopiomelanocortin production, a precursor for the Adrenocorticotropic hormone. If excessively released, it promotes an overproduction of adrenergic neurotransmitters by adrenal glands. Coming at clinical condition known as Cushing syndrome. On the very long term, Nur77 deregulation is known to cause tumors of the pituitary and adrenal glands ((121,122). This happens in pheochromocytoma, which is one of the main VHL disease manifestations. We speculate that continuous VEGF transcription, even in situations where HIF-1α (but not Sp1) is knocked out, could be the explanation for clinical studies where VEGF-targeting drugs have turned out to be effective in kidney cancer treatment as reported in (123). Although we used only confirmed data from the literature, Nur77 may be involved in other regulation systems which were not considered in

our model. The transitions for pVHL fibronectin stabilization show a behaviour which is coherent with biochemical experiments, illustrating a complete abrogation of ECM stabilization and an increased matrix metallo-proteinase action. Although the results are encouraging, the presented model will need further improvements since standard PNs do neither allow a complete transition control nor enzymatic activity modulation. Nevertheless, thanks to its manual curation our model can be used to plan new *in vitro* and *in vivo* experiments. The results are convincing enough to suggest our model as a comprehensive pathway model to simulate the main pVHL functions.

4. Evaluation of the steric impact of flavin adenine dinucleotide in Drosophila melanogaster cryptochrome function.

This chapter has been published in (124) Masiero A, Aufiero S, Minervini G, Moro S, Costa R, Tosatto SCE. Evaluation of the steric impact of flavin adenine dinucleotide in Drosophila melanogaster cryptochrome function. Biochem Biophys Res Commun. 2014 Aug 8;450(4):1606–11.

4.1 Summary

Photoreceptors are crucial components for circadian rhythm entrainment in animals, plants, fungi and cyanobacteria. Cryptochromes (CRYs) are flavin adenine dinucleotide (FAD) containing photoreceptors, and FAD is responsible for signal transduction, in contrast to photolyases where it promotes DNA-damage repair. In this work, we investigated an alternative role for FAD in CRY. We analyzed the *Drosophila melanogaster* CRY crystal structure by means of molecular dynamics, elucidating how this large co-factor within the receptor could be crucial for CRY structural stability. The co-factor appears indeed to improve receptor motility, providing steric hindrance. Moreover, multiple sequence alignments revealed that conserved motifs in the C-terminal tail could be necessary for functional stability. Four systems were tested: (i) dCRY-FAD-Mg^{2+}, (ii) dCRY-Mg^{2+}, (iii) dCRY and (iv) dCRY-FAD-Mg^{2+} with S526-$PO3^{3-}$. All simulations were carried out with NAMD v 2.9, using the CHARMM-27 force field, and were 20 ns long. The multiple sequence alignment was constructed with T-Coffee using a BLOSUM62 matrix. The overall fluctuations of the holoreceptor and phosphorylated system, in absence of light,

suggest that the steric impact of the co-factor inside its binding pocket could be a feature required for receptor activation. Interestingly, the receptor is globally less rigid when the co-factor is contained in its pocket as reported by the RMSF and RMSD plots shown. Thus, we cannot exclude FAD from also having a role due to its molecular structure and hindrance, rather than only being involved in photo reduction and receptor activation itself. The S526 phosphorylation site is close to the CTT, and provides higher fluctuations in this region. Our results suggest phosphorylation could occur before receptor activation, further facilitating receptor activation. This could explain the crucially close CTT position, suggesting a DNA mimicking role for CTT and TIM, result also reinforced by sequence alignment. In this work, I designed and performed the molecular dynamics simulations, implementing the cofactor parameters in the force field, I performed the analyses of the obtained trajectories and the structural comparison between Cryptochrome and (6-4)-Photolyase crystal structures.

4.2 Abstract

Photoreceptors are crucial components for circadian rhythm entrainment in animals, plants, fungi and cyanobacteria. Cryptochromes (CRYs) are flavin adenine dinucleotide (FAD) containing photoreceptors, and FAD is responsible for signal transduction, in contrast to photolyases where it promotes DNA-damage repair. In this work, we investigated an alternative role for FAD in CRY. We analyzed the Drosophila melanogaster CRY crystal structure by means of molecular dynamics, elucidating how this large co-factor within the receptor could be crucial for CRY structural stability. The co-factor appears indeed to improve receptor motility, providing steric hindrance. Moreover, multiple sequence alignments revealed that conserved motifs in the C-terminal tail could be necessary for functional stability.

4.3 Introduction

Light sensing is a fundamental task of living organisms, which have evolved in a rhythmic environment characterized by endless light–dark cycles. Light sensing is present in all kingdoms of life, due to several evolutionary strategies, which have provided organisms with this capability. One of these is light sensing by means of a protein termed photoreceptor (19). Cryptochromes (CRYs) are blue-light sensing receptors, first identified in *Arabidopsis thaliana* in 1993 (23). The hidden nature of their co-factor, which remained unknown for a long time, gave name to this particular class of proteins (19). They are classified as close relatives to photolyases (PLs), with whom they share a major structural part termed photolyase domain and the presence of a FAD co-factor, which is contained in an alpha-helix domain in both PLs and CRYs. PLs work by capturing electrons and providing them to reverse DNA damage, particularly removing pyrimidinic base dimerizations. In the receptor region, where PLs interact with DNA, CRYs show a longer C-Terminal Tail (CTT) (30). In spite of their structural and sequence similarity to photolyases, CRYs show a different kind of activity (31,32). Indeed, they are involved in Insects in circadian rhythm entrainment and magnetosensitivity while in mammals they act as transcription repressors in the main negative feedback loop at the core of the clock. Moreover they lack DNA-repair activity in animals (29). CRYs are therefore widely present in eukaryotic organisms, with a flavin-adenine-dinucleotide (FAD) co-factor (19) providing blue light sensing activity. FAD works as an antenna, capturing photons and transducing them in a chemical reduction response. Unlike mammals, CRY has only one ortholog *Drosophila melanogaster*(dCRY). dCRY is rhythmically expressed in both *Drosophila* clock neurons and the compound eye, responding to light/dark variations (36). During light exposure, dCRY gets activated and opens its CTT(30,32), thus binding to TIMELESS (TIM) and promoting its proteasomal degradation, a process which the F-Box protein Jetlag (Jet) (34,35). In darkness,

TIM forms a dimer with PERIOD (PER). The TIM–PER heterodimer enters the nucleus, inhibiting the CLOCK and CYCLE protein interaction (95,98). On the other hand, when dCRY is binding TIM, it inhibits formation of the TIM-PER heterodimer, allowing CLOCK and CYCLE to interact with a specific E-box DNA segment promoting transcription of clock genes (33,36). This pathway is located at the core of the circadian clock in *D.melanogaster*, and is responsible for its synchronization to the natural light–dark cycles characterized by a 24 h period through dCRY light sensing (36). At a molecular level, receptor activation passes through absorption of a photon by FAD, causing its reduction with electrons towards a radical semiquinone state. FAD chemical reduction appears to play a role in receptor C-terminal tail (CTT) opening. The CTT seems to be involved in protein–protein interactions under light exposure, as it contains plenty of known linear motifs (29), in particular PDZ motifs. One crucial interaction between dCRY CTT and the CTT-coupled motif appears to be mediated by a motif involving the residues FFW on the CTT, as demonstrated in (30). Flavin ring reduction after light exposure and subsequent CTT opening allows TIM, which contains the FFWL motif as well, to interact with dCRY, putatively in the CTT-coupled motif (30). In 2011, Zoltowski and co-workers solved a 2.30 Å resolution X-ray dCRY crystal structure in the dark state (PDB identifier: 4GU5, former 3TVS, Fig. 12) (27). This provided a basis for deeper understanding of the photoreceptor structure and function. FAD is required for receptor activation, as it works like an antenna. However, its structural role in the receptor has not been widely investigated so far.
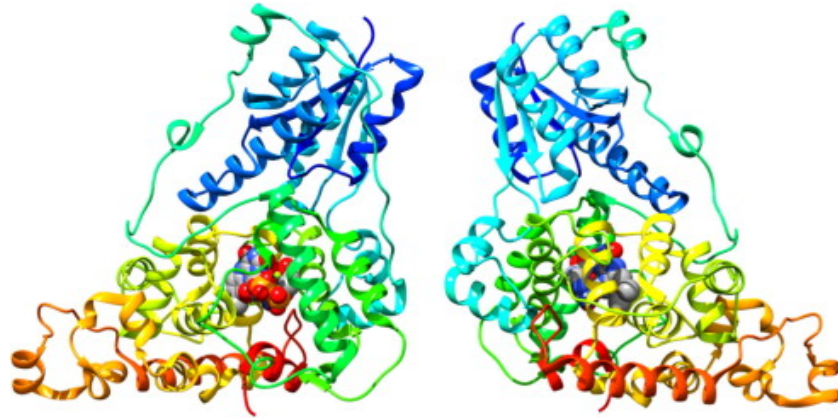
Figure 12: Drosophila melanogaster *cryptochrome structure, PDB identifier: 4GU5. FAD co-factor is represented in spheres, while the receptor is depicted in ribbons.*

To explain the role of the co-factor and of the close Mg^{2+} ion contained in the dCRY FAD binding pocket and their impact on the protein structure, we investigated the time dependent behavior of the holoprotein (HdCRY), the apoprotein (AdCRY) and the system containing only Mg^{2+} (MdCRY) with molecular dynamics (MD) simulations. Moreover, a multiple sequence alignment was performed, adding value and providing new clues on dCRY activation and mechanism. The results of this work suggest that FAD presence has not only a functional meaning, but it may be required to provide an overall increase in fluctuation, decreasing the amount of necessary light input energy to activate the photoreceptor.

4.4 Methods

4.4.1 Molecular dynamics simulations

The crystal structure of dCRY containing two 539 residues long chains was used as starting model for all simulations (PDB code: 4GU5, Fig. 12) (27). The structure contains 30 α-helices (267 residues, 49% of the sequence) and 10 β-strands (37 residues, 6%) and is characterized by the presence of one FAD molecule in bent

conformation (125) and one Mg^{2+} ion in the binding pocket. All simulations were carried out with NAMD v 2.9 (47), using the CHARMM-27 force field (38) on a standard x86 Linux workstation. A phosphorylation on S536 was added through the available NAMD stream files for phosphoserine, which were added to both the topology and parameter files. We used an explicit TIP3p solvent model with cubic boxes of 100 Å × 94 Å × 81 Å, containing ~22,500 water molecules. Periodic boundaries were set at 10 Å from the most external protein atom in the corresponding cartesian axis. All boxes were placed in particle-mesh Ewalds (PME) grids and the overall system charge was neutralized in a 0.150 M NaCl medium. Solvent ions were added with VMD (126), which was also used to generate topology files. The co-factor was kept in its quinonic form in order to avoid any activation-prone movement. The bent conformation (127) was kept during the entire simulations. Oxidized FAD parameters were derived from (128) and implemented in the CHARMM-27 (129). All simulation runs consisted of 100 conjugate gradient minimization steps, 100 ps in NVT conditions, 100 ps in NPT conditions and 20 ns of classical molecular dynamics simulation. In all simulations, the temperature was kept at 300 K and pressure at 1.01325 bar, excluding NVT pre-simulation steps. The integration timestep was 2 fs and the integrator was based on the Verlet method (47). The system was analyzed by four different MD simulations, consisting in the holoprotein crystal structure simulation, a simulation containing a phosphorylation in Ser526 previously reported by Hemsley et al. (29), a simulation containing only the Mg^{2+} ion and a last one without FAD, Mg^{2+} and containing no PTMs. The four runs were then compared in terms of root mean square deviation (RMSD) and root mean square fluctuation (RMSF) to explain the role of each different system component. RMSF indexes are reported as both plots and structure coloring and thickening.

4.4.2 Trajectory analysis

72

All obtained trajectories first went through visual inspection with Chimera (130), cleaning and analysis with Carma (131) and EUCB (132). The first step consisted in water, ion and co-factor removal and system recentering with Carma, obtaining trajectory files containing only the protein structure and a constant fitted center of mass. The second step consisted in backbone RMSD and RMSF index calculation and plot generation with EUCB. In order to obtain information about the flexibility of protein regions in the structures, images were generated coloring and thickening residue ribbons proportionally to their RMSF values with Chimera. To obtain these images, the occupancy column of the structure files was replaced by the residue RMSF value with an *ad hoc* script. All colored structures refer to the last frame of the respective run.

4.4.3 Structure analysis

Both dCRY (27) and (6–4)-Photolyase (PL) (28) structures went through RING(52) and BLUUES (133) analysis. RING generates a network with residues depicted as nodes and interactions depicted as edges, and provides information on conservation and residue interactions. RING found 33 sequences, and calculated overall residue conservation among them. BLUUES provides an electrostatic potential analysis.

4.4.4 Multiple sequence and structure alignments

The previously considered Drosophila-like-CRY sequences (31) were retrieved by Blast search using dCRY as a query. TIM sequences were collected from the same species with the same approach. The multiple sequence alignment was constructed with T-Coffee (134) using a BLOSUM62 matrix. The number of extracted sequences is 36 of which 18 are available for both dCRY and TIM. A sequence logo was built with Weblogo(81)) from this data and overlaid with the experimental dCRY

peptide structure (PDB ID: 4GU5). Intrinsic disorder for dCRY CTT and TIM sequences and secondary structures for TIM was predicted with CSpritz (82). The structures (PDB IDs: 4GU5 and 2WB2) were superposed with Chimera, using the Needleman-Wunsch algorithm with the BLOSUM62 matrix. The RMSD between the protein structures was 0.963 Å.

4.5 Results

4.5.1 Structural impact of the co-factor

The rationale for this work was first the clarification of FAD impact on structure stability by better investigating the behavior of its surrounding. To have a comparison between two FAD containing systems and to better highlight its impact on the structure, we performed a comparison between the holoreceptor, kept as a blank, and the holoreceptor containing a previously reported phosphorylation in S526 (PdCRY), which is a known occurring event (29)). In order to investigate the co-factor impact on its surrounding protein elements, MD simulation outputs were evaluated in terms of RMSD and RMSF. *A priori*, we hypothesized that FAD presence would improve the overall receptor stability through intermolecular interactions and its removal from the binding pocket would yield receptor misfolding. To evaluate the structural role of the co-factor, the first index used was RMSD from trajectory files obtained by 20 ns MD simulations. RMSD evaluates the mean receptor backbone distance between the first and other frames generated for a given trajectory. RMSD plots of the systems showed how the more stable conformations are those not containing the co-factor (see SM 2, Appendix). Very recently, Vaidya and co-workers demonstrated *in vitro* that the structural elements surrounding the CTT are exposed to Trypsin cleavage under light conditions, suggesting disorder (30). They also report a general photoreceptor structure involvement after light exposure. Our study focuses on the effect of FAD itself

74

before the reaction, considering its effect on the overall structure. HdCRY RMSD plot reveals a mean plateau value of 4.5 Å, and the phosphorylated system an even higher peak of almost 5.0 Å. In the latter system, the RMSD slope is much higher compared to other systems (see SM 2, Appendix), highlighting the strong impact of the post-translational modification on the system. RMSD values of the other two systems containing no FAD and neither FAD nor Mg^{2+} are lower, suggesting more rigid behaviors. The second output analysis was RMSF, measuring the mean backbone RMSD value for each residue during an entire run, distinguishing more fluctuating and more rigid regions of a system. The RMSF plots showed that the overall structures not containing the FAD co-factor are more rigid and the fluctuations minimized (see SM 3, Appendix).
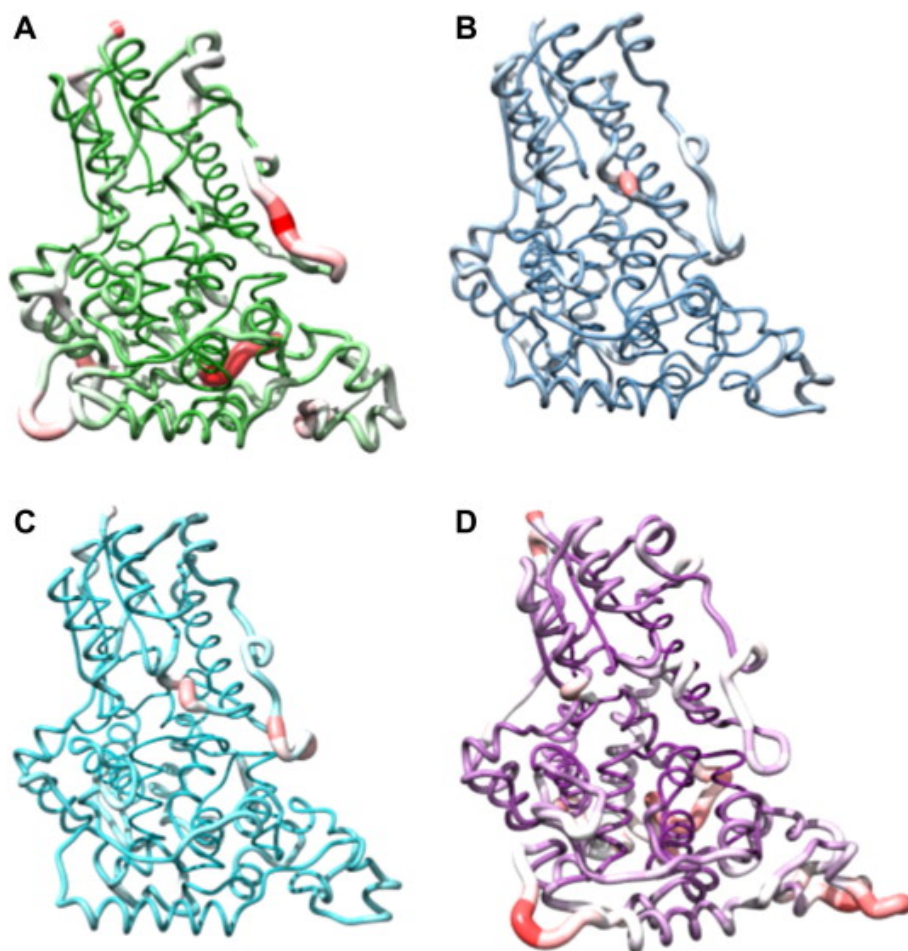
Figure 13: RMSF thickening and coloring for (A) holoprotein, (B) cryptochrome without FAD, (C) cryptochrome without both FAD and Mg^{2+}, (D) cryptochrome with phosphorylated S526.

The regions more affected by FAD presence, according to RMSF plots, are residues 200–300 and the CTT region. The first corresponds to the FAD phosphate binding pocket surface moiety boundaries, a solvent exposed segment and the protrusion motif (see SM 1, Appendix). A comparison between HdCRY, AdCRY and MdCRY dynamics clearly shows that the holoreceptor region between residues 200 and 300 is fluctuating with peaks reaching 4.0 Å, while in AdCRY and MdCRY these values are lower at 3.5 Å and 2.5 Å respectively (see SM 3, Appendix). Another

region deeply biased by FAD presence/absence is the so-called CTT base loop (residues 154–160) (30). The highest motility was found in the system containing a phosphorylation in S526, the Mg^{2+} ion and the FAD co-factor. The phosphoserine forms hydrogen bonds with both the Phe428 and Glu429 backbone (C-Terminal lid) for almost all analyzed frames. The RMSF plot of this system shows slightly lower peaks in the 200–300 region, with values reaching 3.5 Å, but an improved CTT motility compared to all other RMSF plots (see SM 3, Appendix). These results show that FAD presence in the receptor improves overall motility, as well as the kinetic energy of the system. AdCRY and MdCRY were definitely more stable and rigid (see SM 3, Appendix, and Fig. 13). This can be explained with steric effects and is clearly noticeable by RMSD plot comparison (see SM 2, Appendix). The slope of the RMSD plot for systems without FAD is dramatically lower. Interestingly, RING network analysis revealed that residues with more than 80% conservation (87 residues) are mostly kept in the hydrophobic protein core (see SM 1, Appendix). Of those 87 residues, 16 are in close contact with the cofactor. This analysis was complemented with a multiple sequence alignment, which also revealed high conservation also for some CTT residues. The conservation rate decreases proportionally to Phylum distance. BLUUES electrostatic potential map comparisons between dCRY and (6–4)-PL revealed a positive potential in both dCRY and (6–4)-PL, (see SM 4, Appendix). In this region, (6–4)-PL contains positively charged residues interacting with DNA. In dCRY, this region corresponds to the CTT lid, which remained rigid during the simulation. This result, in addition to structural superposition, could suggest a DNA mimicking role of this region.

4.5.2 The FFW motif and S526 phosphorylation effect

S526 superposes to a DNA phosphate moiety in the *D.melanogaster* PL crystal structure. The same DNA phosphate moiety was found to partially superpose with dCRY E530 as well, suggesting a DNA-mimicking role of the CTT.
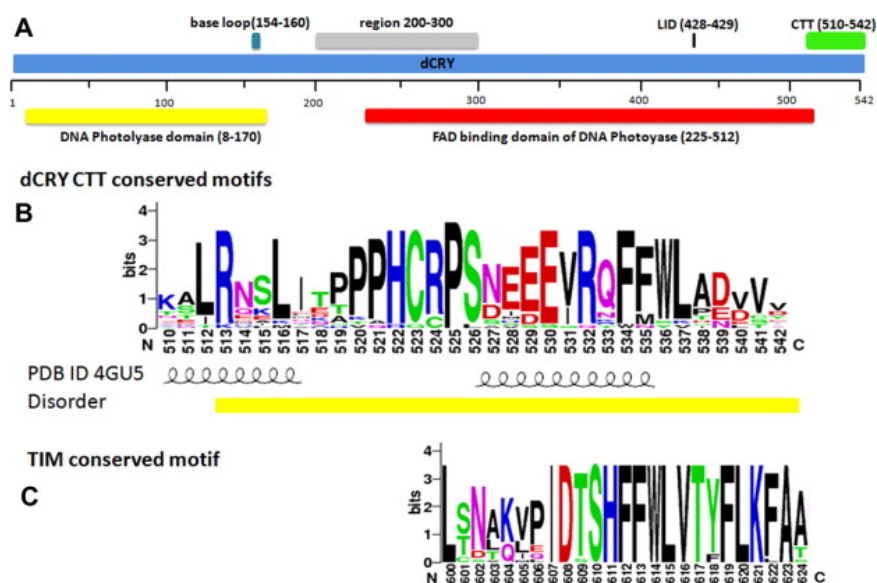
Figure 14: (A) Schematic representation of dCRY, below the regions Photolyase-like, above the most important regions influenced by presence/absence of FAD co-factor. (B) Overview of the dCRY CTT features. dCRY CTT sequence logo shows the conserved motifs above, features from the crystallographic dCRY structure (PDB identifier 4GU5) and predicted disorder are showed below. The secondary structure is shown as helix. Predicted disorder is shown as a yellow line. (C) TIM sequence logo shows the FFW conserved motif suggested to interact with dCRY receptor region, secondary structures (S.S.) prediction and structured region are showed below. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Furthermore, the mainly aromatic ring nature of the residues of the FFWL motif is suggesting some similarity with nucleic acid base rings. The importance of these motif residues can find partial confirmation in the multiple sequence alignment, which revealed a high conservation of both the SNEEE and FFW motifs through all analyzed Drosophila-like-CRYs (see Fig. 14). S526 is 100% conserved, while the FFW motif the residues are 100%, 88%, 100% conserved, respectively (see Fig. 13). The sequence logo shows how the CTT linear motifs are located in a segment with strong sequence conservation, in a putative α-helix inside an intrinsically disordered region (see Fig. 13).

4.5.3 TIM sequence analysis

Starting from the reported FFW motif correspondence between TIM and dCRY CTT (30), we extended the analysis to TIM sequences belonging to the same

species. We found that TIM keeps the FFW motif highly conserved (100%) among all explored species (Fig. 14). This suggests that the conserved tryptophan of TIM and the dCRY CTT FFW motif could interact in the same manner with dCRY. Conversely, neither motif was found among *D.melanogaster* Photolyase orthologs.

4.6 Discussion

Light signal transduction into chemical responses is frequent in nature. Here we investigated whether the presence of FAD could have not only a light sensing role within photoreceptors, but could be also required to provide a more suitable overall structure to maximize receptor reactivity towards light impulses. The overall fluctuations of the holoreceptor and phosphorylated system, in absence of light, suggest that the steric impact of the co-factor inside its binding pocket could be a feature required for receptor activation. Interestingly, the receptor is globally less rigid when the co-factor is contained in its pocket as reported by the RMSF and RMSD plots shown. Thus, we cannot exclude FAD from also having a role due to its molecular structure and hindrance, rather than only being involved in photo reduction and receptor activation itself. This could be explained considering the more fluctuation-prone HdCRY behavior, likely requiring less input energy to go towards a state transition and activation. The RMSF value of the CTT preceding region is reinforcing this hypothesis. All systems containing FAD have higher RMSF values towards the CTT between residues 450 and 525. Whenever FAD is lacking, this region is more rigid and less ready to move. Given that receptor activation appears to be dependent on this residue segment, we can speculate that FAD could also have a prominent role before receptor activation unconnected to its intrinsic light exposure reactivity. Analysis of the phosphorylated system (P-dCRY) revealed an even higher overall fluctuation compared to all other systems. The S526 phosphorylation site is close to the CTT, and provides higher fluctuations in this region. Our results suggest phosphorylation could occur before receptor activation, further facilitating receptor activation. This could explain the crucially close CTT

position, suggesting a DNA mimicking role for CTT and TIM. Furthermore, dCRY E530 and S526 are close to a PL phosphate DNA moiety after superposition (0.98 Å and 2.25 Å, respectively for the closest atoms), thus reinforcing what previously demonstrated in (135). On the other hand, multiple sequence and structural alignments also revealed that the CTT could mimic the damaged DNA Photolyase-DNA interaction and establish contacts with the FAD binding pocket. In particular, the aromatic rings of the FFW motif residues show similarity with nucleic acid base rings and are conserved in the main dCRY interactor TIM (30). In the end, our data suggest that even if chemical reactivity of the co-factor is the main reason behind receptor activation, steric hindrance of the molecule could be a required feature as well. The multiple sequence alignment also showed that the SNEEE and FFW motifs are conserved in CRY ortholog CTTs. These conserved species-specific motifs, which are not present in the Photolyase family, could suggest a functional role in the photoreceptor. Light exposure induces dCRY activation, opening its CTT (29,30) and allowing binding to TIM, which it is subsequently degraded by the proteasome (34,35). Further *in vitro* validation of the obtained results should be carried out in order to shed light on these issues, due to the *in silico* origin of data. Our results nevertheless suggest that presence of the conserved FFW motif in both TIM and dCRY and SNEEE, once phosphorylated, and FFW motifs in dCRY could mimic DNA interaction with the CTT coupled motif. Finally, these findings are suggesting an additional prominent structural role of the co-factor in the receptor activation process, providing another small piece to the enormous light stimuli response mechanism jigsaw.

5. Structural protein reorganization and fold emergence investigated through amino acid sequence permutations

5.1 Summary

Correlation between random amino acid sequences and protein folds suggests that proteins autonomously evolved the most stable folds, with stability and function evolving subsequently, suggesting the existence of common protein ancestors from which all modern proteins evolved. To test this hypothesis, we shuffled the sequences of 10 natural proteins and obtained 40 different and apparently unrelated folds. Our results suggest that shuffled sequences are sufficiently stable and may act as a basis to evolve functional proteins. The common secondary structure of modern proteins is well represented by a small set of permuted sequences, which also show the emergence of intrinsic disorder and aggregation-prone stretches of the polypeptide chain. Sequences were permuted following this crossing scheme: ABCA'B'C' becoming A'AB'BC'C and so on. Fold prediction was carried out with Rosetta ab initio software, and all 4 ns long molecular dynamics simulations of all the proteins were carried out with GROMACS, with CHARMM-27 all atoms force-field. The disorder predictions, aggregation predictions and secondary structure predictions were carried out with PASTA2.0 software. Our opinion is that the modern folds may result from biological evolution operating in the form of continuous finishing of a finite set of randomly generated proto-folds.

We demonstrated that shuffled sequences might act as a basis to evolve functional proteins, with the common secondary structure of modern proteins well represented by a small set of permuted sequences. Although meaningful, the results presented here are theoretical and their reliability is related to the validity of the underlying in silico predictions. In this work, I performed the analysis of the molecular dynamics simulations trajectories of the predictions, and I performed the disorder, electrostatic, aggregation and secondary structure calculations with PASTA 2.0.

5.2 Abstract

Correlation between random amino acid sequences and protein folds suggests that proteins autonomously evolved the most stable folds, with stability and function evolving subsequently, suggesting the existence of common protein ancestors from which all modern proteins evolved. To test this hypothesis, we shuffled the sequences of 10 natural proteins and obtained 40 different and apparently unrelated folds. Our results suggest that shuffled sequences are sufficiently stable and may act as a basis to evolve functional proteins. The common secondary structure of modern proteins is well represented by a small set of permuted sequences, which also show the emergence of intrinsic disorder and aggregation-prone stretches of the polypeptide chain.

5.3 Introduction

Proteins are considered as the building blocks of life owing to their stability and functional plasticity. At the molecular level, proteins are heteropolymers of 20 different amino acids. Simple dipeptides such as Gly-Gly or Ser-His were reported to catalyze both proteolytic and peptide bond formation (115,116). It is

conceivable that small catalytic proto-enzymes contributed to the evolution of longer copolymers, triggering the race for the most stable polypeptide chain. It is also conceivable that a stable chain incorporated the same catalytic motifs, perhaps yielding a first complex enzyme (e.g., a catalytic center coupled with a stable scaffold). Over time proteins evolved different and complex functions such as enzymatic catalysis, structural, and signaling functions. Current theories regarding the origin of life suggest that amino acid chains co-evolved with, or immediately after, the emergence of autocatalytic RNA (139,140). On the basis of these theories, primordial proteins, also known as proteinoids, promoted the nucleic acid onset, allowing their stabilization and conferring the ability to adapt to environmental changes(141)). Ideally, primordial RNA–polypeptide association could have enlarged the spectrum of different catalyses allowing the formation of more complex reactions. Notably, in modern organisms peptide bond formation is driven by the ribosome, a highly efficient RNA–protein complex. Considering a prebiotic scenario where rigid environmental conditions coupled with an unstable chemical surrounding promoted protein degradation, a stable structure could be considered as a favorable requirement for protein evolution. Indeed, since the appearance of proteins, only few folds became fixed, perhaps as a result of thermodynamic (or biological *ante litteram*?) selection of the most stable three-dimensional structures. Correlations between random amino acid sequences and protein folds were investigated by Weiss et al.(142), suggesting the general idea that proteins autonomously evolved the most stable folds only after sequence evolution "refined" the surviving folds in order to acquire stability and function (143,144). Indeed, modern proteins are characterized by a large number of different sequences not balanced by the same number of different fold(145)). Unexpectedly, structural analysis in silico on random proteins suggested that polypeptide chains may fold and adopt conformations comparable to natural proteins (37). Proteinoid theory assumes that simple enzymatic function evolved spontaneously from simple and not biologically generated polypeptides (146). On the other hand, it is credible that

the first proteins were generated through ligation of small and stable "amino acid modules" operated by proto-enzymes characterized by both low specificity and efficiency. In this scenario, small proto-proteins should be responsible for all resulting modern folds. In this work, we investigated the occurrence of different folds using an in silico combinatorial approach. We shuffled the sequences of 10 natural proteins and obtained 40 unrelated different folds. A large spectrum of different organizations were observed, including all-α, β-barrel, and disorder. Finally, our results suggest that abiotic complex fold emergence may have resulted from stochastic sequence rearrangement.

5.4 Methods

Permutations were manually performed starting from 10 different sequences in Fasta format of the same number of natural proteins selected from the Protein Data Bank (PDB codes: 1ARK, 1HX2, 1KV0, 1PPT, 1TCP, 1UOY, 1ZFI, 2BHI, 2CDX, 2KJF) (39). In order to minimize both the error and computational demands associated with the ab initio methodology, we selected proteins with a maximum length of 66 amino acids.

```
>1ARK:A|PDBID|CHAIN|SEQUENCE
TAGKIFRAMYDYMAADADEVSFKDGDAIINVQAIDEGWMYGTVQRTGRTGMLPANYVEAI


TAG KIF RAM YDY MAA DAD EVS FKD GDA IIN    VQA IDE GWM YGT VQR TGR TGM LPA NYV EAI
 1   2   3   4   5   6   7   8   9  10      11  12  13  14  15  16  17  18  19  20


>PERMUTATION 1 1ARK
VQA TAG IDE KIF GWM RAM YGT YDY VQR MAA    TGR DAD TGM EVS LPA FKD NYV GDA EAI IIN
 11  1  12   2  13   3  14   4  15   5      16   6  17   7  18   8  19   9  20  10

>PERMUTATION 2 1ARK
TGR VQA DAD TAG TGM IDE EVS KIF LPA GWM    FKD RAM NYV YGT GDA YDY EAI VQR IIN MAA
 16  11   6   1  17  12   7   2  18  13       8   3  19  14   9   4  20  15  10   5

>PERMUTATION 3 1ARK
FKD TGR RAM VQA NYV DAD YGT TAG GDA TGM    YDY IDE EAI EVS VQR KIF IIN LPA MAA GWM
  8  16   3  11  19   6  14   1   9  17       4  12  20   7  15   2  10  18   5  13

>PERMUTATION 4 1ARK
YDY FKD IDE TGR EAI RAM EVS VQA VQR NYV    KIF DAD IIN YGT LPA TAG MAA GDA GWM TGM
  4   8  12  16  20   3   7  11  15  19       2   6  10  14  18   1   5   9  13  17
```

Figure 15: Shuffling schema. Each sequence was initially split into 20 blocks of three amino acids long and then divided in two groups termed A and B. The first block of group B was placed at the beginning of the new permuted sequence and the first block of A was then placed second. The same was applied for the other blocks until a new permuted sequence was obtained.

Each sequence was initially split into 20 blocks of three amino acids long and then divided into two halves, termed A and B. The first block of group B was placed at the beginning of the new permuted sequence and the first block of A was then placed second. The same was applied for the other blocks until a new permuted sequence was obtained. A schematic explanation of the permutation schema is shown in Fig. 15. The three-dimensional model structures of both natural and permuted proteins were predicted using Rosetta (147), a piece of ab initio protein structure prediction software based on the assumption that local interactions bias the conformation of sequence fragments in a polypeptide chain, while global interactions determine the three-dimensional structure with minimal energy (147). For each sequence, 25,000 decoys were generated and clustered using the integrated clustering module. Only the best ranked model proposed for each sequence was considered. The predicted three-dimensional structures of native sequences were calculated and used to test the setup applied in the ab initio protocol. A comparison between experimental and predicted structures is shown in Supplementary Fig. 1. The overall fold stability was studied by performing 4 ns of

molecular dynamics (MD) simulation with Gromacs (46) using the CHARMM 27 force field (38). At 4 ns per system the overall simulation time for all predictions was 200 ns. The total energy of the protein in solution was calculated with BLUUES (133). Disorder was predicted using ESpritz (148) and both secondary structure and aggregation propensity with PASTA 2.0 (129).

5.5 Results

The aim of the present work was to elucidate the way evolution generated the variability observed among existing protein folds. In this sense, we hypothesized a credible scenario where several different sequences arose from a common mixture of short prebiotic polypeptide chains (e.g., a primordial broth). Starting from an existing protein, we split the sequence into smaller segments of three amino acids each, then reassembled the blocks to obtain four new shuffled sequences. The resulting sequences conserved the same overall amino acid composition as the natural one, but different internal organization, as shown in Fig. 16.

5.5.1 Fold analysis

The propensity of shuffled proteins to generate stable folds was measured by calculating a three-dimensional structure for each permuted sequence with Rosetta (147). Interpretations of the results described here are heavily related to the validity of structures predicted using the ab initio method. In a number of cases, Rosetta was shown to perform fairly well and even produced near-atomic resolution structures (150,151). Our results for the natural proteins used in this research confirm that the predictions are in most cases accurate in terms of overall fold, secondary structure content, and topology (Supplementary Fig. 1(136))). The same protocol was also applied to the shuffled sequences. The structures obtained show that pseudo-proteins tend to assume a well-ordered three-dimensional

structure, with almost all predictions promoting a compact fold (Fig. 16). Notably, in several tests the structure obtained at permutation four resulted in a completely different secondary structure content with respect to the starting structure. This finding reinforces the idea that the propensity to evolve different folds is an emergent property of amino acid chains and clearly showed that a sequence can freely evolve different three-dimensional shapes when not constrained by the requirements of biological evolution (37). An illustration of this evidence may be given by the permutation of carnocyclin (PDB code 2KJF) (Fig. 16). The native protein folds in an all-α structure and is known to be very stable, with a fold shared with other proteins of the bacteriocin family (152). Owing to its important biological function, the general fold of this protein family is largely conserved between bacteria. When permuted, the protein gave a wide spectrum of different organizations.
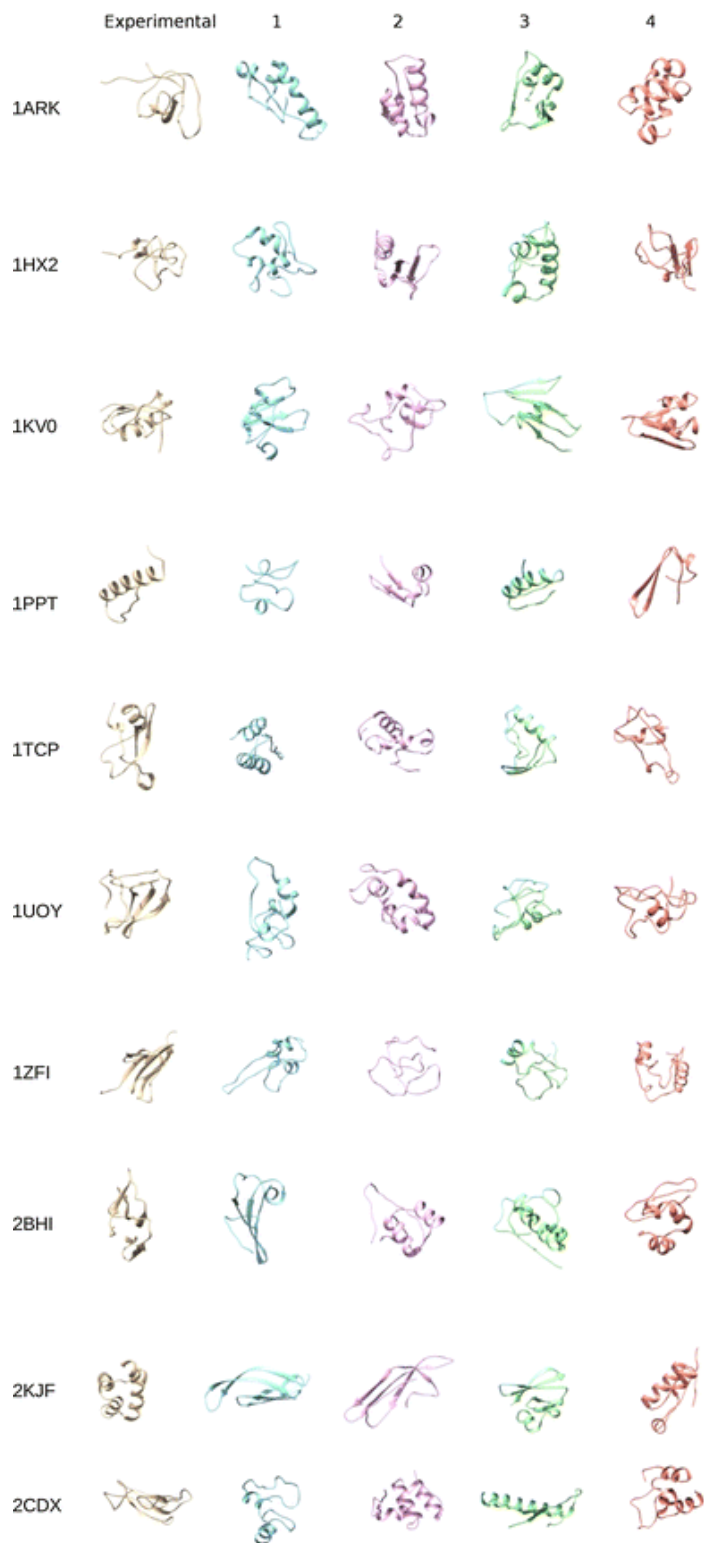
Figure 16: Prediction of permuted proteins. Comparison between experimental structure and predicted structures after permutation. Brown represents the experimental structure; blue, purple, green, and orange represent permutations 1, 2, 3, and 4, respectively. The corresponding PDB code is reported on the left

Indeed, its permutation generates a β-barrel-like, plain β-sheet, α–β mixed, and finally an all-α structure, at permutation steps 1, 2, 3, and 4, respectively. Similar permutation behavior was obtained for a trypsin inhibitor (PDB 1HX2) and avian pancreatic polypeptide (PDB 1PPT). More generally, in all simulations we observed an alternation between several folds with the protein structure suddenly changing between permutations. The result confirms in part what was previously reported by Luisi and co-workers(153), i.e., the identification of folded proteins emerging from a random library. On the other hand, variability may suggest that different folds can spontaneously derive from a small sequence subset as obtained from only four permutation events.

5.5.2 Stability analysis

Every predicted protein, native or shuffled, went through a short (4 ns) MD simulation with the aim of evaluating the stability of the generated folds in aqueous solution. Counting every generated sequence and prediction, 50 MD simulations were carried out and evaluated through both visual inspection and root mean square deviation (RMSD) plots. This helped direct visual detection of the most widely fluctuating systems. For the majority of the targets this revealed no wide nor fast misfolding behavior (Supplementary Fig. 2 (136)). The plots of the shuffled predictions in some cases show slightly higher values than the predictions, which is probably due to structural relaxation and energy minimization. This behavior is also confirmed by the total BLUUES (133) energy (see Fig. 17). A clear example can be seen in carnocyclin (PDB code 2KJF), where RMSD variation of the shuffled proteins reaches 500 % the native prediction, the highest overall value. This can be partially explained by the non-shuffled prediction RMSD plot, which has a rigid overall structure. This protein also shows the highest variability among different folds. However, no misfolding was observed during the simulations. Although the simulation time for each model was kept short because of the large number of systems to test, no misfolding events were observed and in almost all cases RMSD

values were lower than 0.4 nm. Considering this result we can assume that permuted polypeptide chains show a structural stability comparable to the natural proteins.

5.5.3 Homology search and disorder analysis

In order to evaluate the effect of sequence similarity on the permuted sequences, we performed a similarity search with Blast (154). The native sequences were all identified, but the search for permuted sequences gave no significant results. In other words, this finding confirmed that the tested sequences should be considered non-natural. Indeed, we were confident that structural predictions were not (or minimally) influenced by homology with other known proteins. The permutation of leech carboxypeptidase inhibitor (PDB 1ZFI) presents another peculiar result, showing a disordered fold for permutation 2 (Fig. 16). Similar results were also obtained for both avian pancreatic polypeptide (PDB 1PPT) and the so-called bubble protein (PDB 1UOY) at permutation 1. Intrinsically disordered proteins are biological entities showing important biological activities (i.e., protein–protein interaction) characterized by existing in a constitutively unfolded state (155). As Rosetta was not originally developed to predict disorder, we tested the sequences with ESpritz (148) to confirm this finding (see Fig. 17). The analysis revealed that only few pseudo-proteins tend to assume a disordered state as the main secondary structure organization.

	KJ/mol
1ARK	-15572.162
permutation 1	-15572.894
permutation 2	-15409.516
permutation 3	-15465.757
permutation 4	-15570.566
1HX2	-14315.641
permutation 1	-14223.167
permutation 2	-14676.260
permutation 3	-14455.909
permutation 4	-14260.513
1KV0	-16292.801
permutation 1	-16136.463
permutation 2	-16042.270
permutation 3	-16146.503
permutation 4	-16103.350
1PPT	-10514.172
permutation 1	-10498.864
permutation 2	-10451.719
permutation 3	-105311.903
permutation 4	-9893.070
1TCP	-17029.855
permutation 1	-17112.712
permutation 2	-17132.655
permutation 3	-17123.834
permutation 4	-17030.045
1UOY	-17384.146
permutation 1	-17207.211
permutation 2	-17308.351
permutation 3	-17142.155
permutation 4	-17247.432
1ZFI	-16295.205
permutation 1	-15798.162
permutation 2	-15903.024
permutation 3	-16045.333
permutation 4	-16201.315
2BHI	-26887.767
permutation 1	-13323.569
permutation 2	-13291.710
permutation 3	-13396.177
permutation 4	-13323.412
2CDX	-13558.024
permutation 1	-15231.255
permutation 2	-13919.995
permutation 3	-13771.731
permutation 4	-13932.128
2KJF	-12381.117
permutation 1	-12100.482
permutation 2	-12017.432
permutation 3	-12225.131
permutation 4	-12164.205

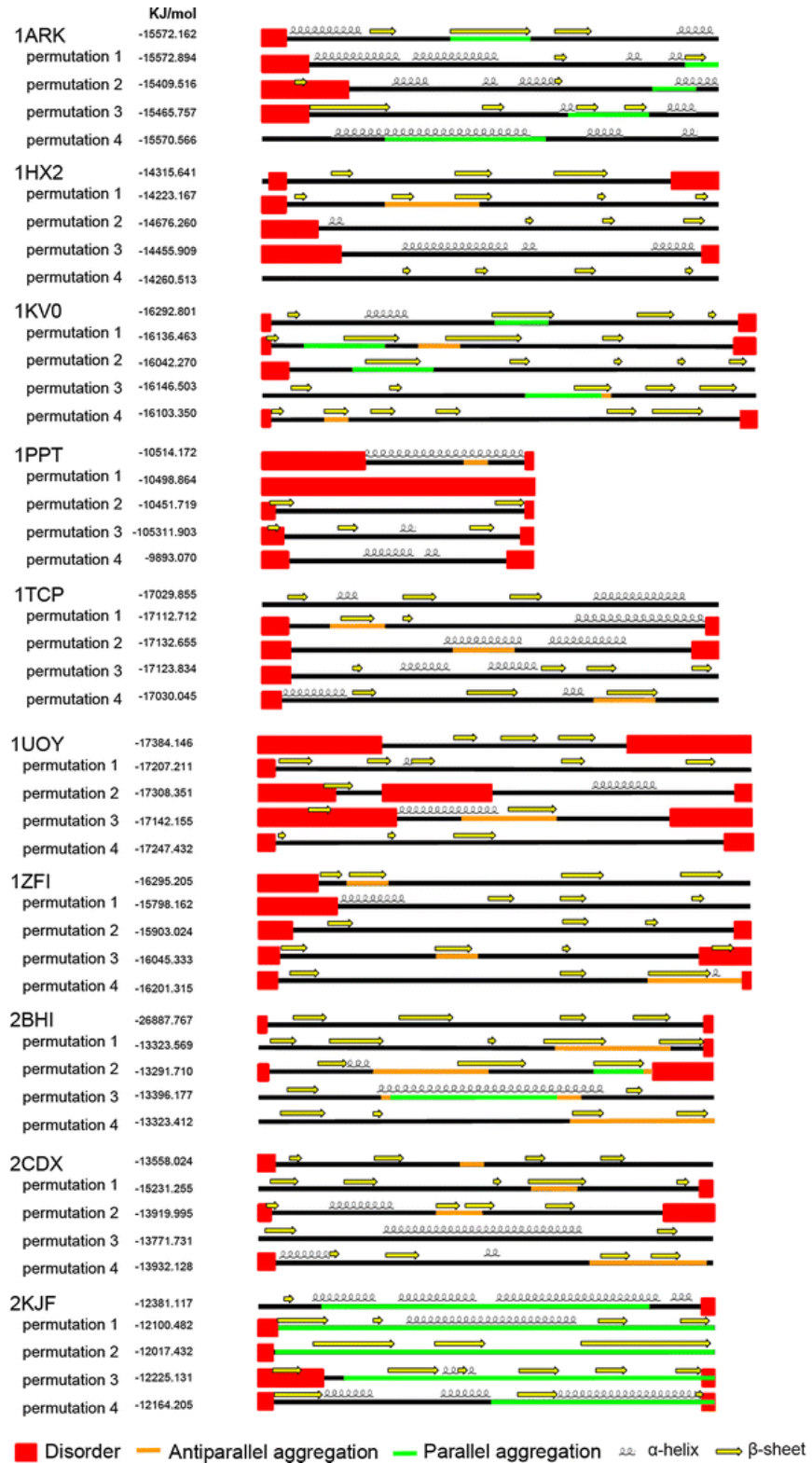■ Disorder — Antiparallel aggregation — Parallel aggregation ‍ α-helix ⟹ β-sheet

Figure 17: Sequence features and energy of the permuted proteins. A sequence diagram is shown for each experimental protein and its four permutations, highlighting predicted secondary structure (spirals for helices and arrows for strands), disorder (thick red line), and aggregation propensity (narrow yellow and green lines) according to the symbols explained at the bottom of the figure. The total BLUUES energy of the protein structure is shown after the sequence identifier.

91

The average disorder content (17.1 %) is slightly higher than the wild-type one (16 %), with the notable exception of PDB code 1PPT. A similar but stronger trend is also observed for aggregation propensities measured with PASTA 2.0 (149), where the overall percentage of aggregating residues rises from 11.45 to 20.16 %, suggesting a negative selection for aggregating sequences. Although consistent with structure prediction, this finding should be considered only as preliminary, as low sequence similarity may induce the disorder and aggregation predictors to produce false positive output. Considering all results in a prebiotic context, we can speculate that the major structural organization observed in modern natural proteins can be evolved from a simple rearrangement of the same building blocks.

5.6 Discussion

Shared fold and sequence similarity are common traits of modern proteins as shown in, e.g., the SCOP database (136). Current accepted theories explain this finding assuming a hypothetical common ancestor from which all modern organisms derived (157,158). In other words, all current folds could have been fixed when life emerged on Earth and survived until today owing to their intrinsic stability. In this work, we explored one of the possible solutions adopted by nature to generate different folds from prebiotic-compatible building blocks. We shuffled the sequences of natural proteins in order to obtain new permutated proteins. In 2002, spontaneous formation of a chain of 10–15 amino acids long in rigid prebiotic conditions was reported(159). The authors suggested that evolution of large proteins is compatible with a fickle environment, at least chemically speaking. In this sense, we hypothesized a scenario where different proto-proteins arose from a mixture of short prebiotic polypeptides. In previous work, Kauffman(160) argued that the ability to act autonomously in an environment is a fundamental characteristic of life. In other words, both diverse and stable folds provide the stability needed for effective biological evolution. In 2006, Luisi and co-workers

reported that random small polypeptides spontaneously tend to assume both a compact and stable conformation (153). It is useful to recall that modern proteins share only a limited number of folds, generally combined to form multi-domain architectures(145). One can argue that this is due to the biological evolution, with modern folds representing the best solution found by the nature to solve a specific problem. Our opinion is that the modern folds may result from biological evolution operating in the form of continuous finishing of a finite set of randomly generated proto-folds. We demonstrated that shuffled sequences may act as a basis to evolve functional proteins, with the common secondary structure of modern proteins well represented by a small set of permuted sequences. Although meaningful, the results presented here are theoretical and their reliability is related to the validity of the underlying in silico predictions. Previous work suggested that folding is an innate property of amino acid chains(143,153,161)). Nevertheless, we found several sequences apparently promoting disorder and characterized by low complexity. During the last decade, disorder assumed relevance in protein science with a large number of intrinsically unstructured proteins discovered to be involved in fundamental cellular processes such as protein–protein interactions, scaffolding, signaling, and transcription (155). Our research suggested that both structured and intrinsically unfolded proteins can evolve by means of simple amino acid recombination. This finding, if experimentally validated, will perhaps provide a better understanding of the driving forces favoring the emergence of Life.

6. RING MD: gathering time into structures

6.1 Summary

Several methods have been developed and tested for molecular dynamics (MD) simulations output analysis. Most of them focus on Root Mean Square Deviation (RMSD) and Root Mean Square Fluctuation (RMSF). The main idea of this work is to provide the scientific community a method that directly analyses a MD simulation trajectory file. The method performs RMSD and RMSF calculations, as other methods do. It will also provide per-residue calculation of H-bonds, salt bridges, disulphide bridges, π-π stacks, cation-π interactions, van der Waals contacts and a comprehensive interaction calculation of the frequencies of occurrence of these interactions. The method is based on RING networks. A network is generated for every given frame of the trajectory, and extract information that can discriminate key events of the MD simulation. Finally, the user is also provided of a PDB structure with residues coloured proportionally to the frequency of occurrence of the selected type of interaction, and of a contact map highlighting the most important interactions responsible for protein structural maintenance. The comparison between classical MD analysis and RING MD analysis has been performed on three well studied targets. The first one, human erythrocitic ubiquitin is 76 residues long and is composed by 23% of α-helices and 34% of β-sheets. The second one, bacteriophage T4 lysozime, is 164 residues long and is composed by 66% of α-helices and 9% of β-sheets. The third target, bacteriophage T4 glutaredoxin, is an NMR solution structures ensemble, is 87 residues long and is composed by 35% of α-helices and 22% of β-sheets. In the latter case, we chose the first model in the PDB file for the simulations. All MD simulations were carried out with NAMD v 2.9, using CHARMM-27 force field. All clusterings were carried out with kmeans algorithm. The results of proteins analyses are: (i) Ubiquitin: Considering all the computed interaction and all the obtained clusters, the key

residues involved in conformational transitions for Ubiquitin appear to be D39, Q40, R42, R72 and L73. These residues appear to be involved in every clustering as driving conformational changes residues, moreover R72 has been proven to be target for Insulin Degrading Enzyme (IDE) and the difference between clusters is providing insights for the activity of this enzyme simply from strcture. (ii) Lysozyme: At a more comprehensive level, considering all the interactions, the key residues driving conformational changes appear to be E22, G30, E45, D47, K48, V87, D92, N101, K124, R137 and D159. (iii) Glutaredoxin: a comprehensive evaluation of the whole set of the considered interaction lead to identify Y5, D18, R22, E32, N35, F43, T64, Q67, D72, S74, D80 and Y85 as key residues for overall conformational transitions during the entire 50 ns MD simulation. The capability of this work-flow to speed up the analysis of an MD trajectory, could turn out to be very useful to the scientific community facing more and more MD simulation experiments. As an example, the parameters obtained from the clustering can be interpreted as the parameters of a mixture model. Therefore, one can use such model to test if a new MD is generated from the same stochastic process. A clear instance of this idea is the evaluation of mutations impact in the protein dynamic or in its stability. There are plenty of cases where the recognition of protein conformers can be of interest, and RING MD clearly establish a solid framework for a sound MD analysis. As a future work, we plan to improve conformers assigned to clusters. In this work, I provided the idea for the design of the tool and performed all the simulations, relating the results with biological evidence I found in the literature. I also performed the comparison between the classical and the proposed MD trajectory analyses.

6.2 Abstract

Molecular Dynamics (MD) simulations gained increasing relevance over the last years, but the complexity related to their interpretation and analysis is still one of the major challenges for most users. In this work we introduce RING MD, a tool for

the analyses and interpretation of MD simulations. RING MD is able to identify the most important frames (PDB structures) and key residues that cause different conformers to be observed, providing a simple interpretation useful for non-expert users. Comparison with the classical analysis of three MD simulations confirmed RING MD results and effectiveness.

6.3 Introduction and methods

Molecular dynamics (MD) is nowadays a key toolbox for disease studies (54,162–164) and protein characterization (124,165–167). MD gained further relevance after a Nobel Prize award for Chemistry was assigned for insights in this field ((38). Despite its proven effectiveness, the knowledge required to interpret and gather information from classical MD output still represents the greatest issue for the scientific community. RING MD is a tool designed to simplify the analysis and interpretation of MD experiments, which are otherwise dependent on human expertise. Networks ((52,168) are increasingly used in biology, but little work was done to analyze MD simulations ((169,170). Residue Interaction Networks (RINs) forms the basis of this work, thanks to their proven ability in the representation of key protein features and contacts (52,168). Every MD simulation can be seen as multiple snapshots of the same structure, changing its conformation and fluctuating around the so-called equilibrium state. As the conformation changes, the respective RIN changes as well. The RING web server computes a RIN where 6 different chemical interactions are identified, namely hydrogen bonds, π-cation interactions, π-π stacks, ionic bonds, disulfide bonds and van-der-Waals interactions. Generating a network from a single experimental PDB is nevertheless a rough approximation of the true RIN, as the stochastic molecular fluctuations can modify the edges in the protein network. Ideally, RIN interactions should be extended to highlight (i) *possible* or (ii) *time-dependent* edges (see Fig. 18). We performed RING MD analysis on 50ns MD simulations(47,126,131,132)of three

well-known targets, Ubiquitin (171), Glutaredoxin(172)and T4 Lysozyme (173)to validate the tool (see Appendix). For each snapshot of the simulation, RING MD computes different conformer networks called Time-Dependent Contact-Maps (TDCM) (169,174,175), which describes the *possible* states explored during time. These networks are different, as the intrinsic protein flexibility highlights different protein states. To address the high sparcity of amino acid interactions, we encoded RINs to a set of binary values associated with each residue. Amino acids involved in a chemical bond are the only active variables (see Fig 1 and online methods). We captured protein variability using k-means clustering algorithm (174), which grouped RINs into k similar protein conformers that model part of the MD. The amount of clusters k is a tradeoff between model likelihood and number of parameters as defined by the Bayesian information criterion (see Appendix for details).
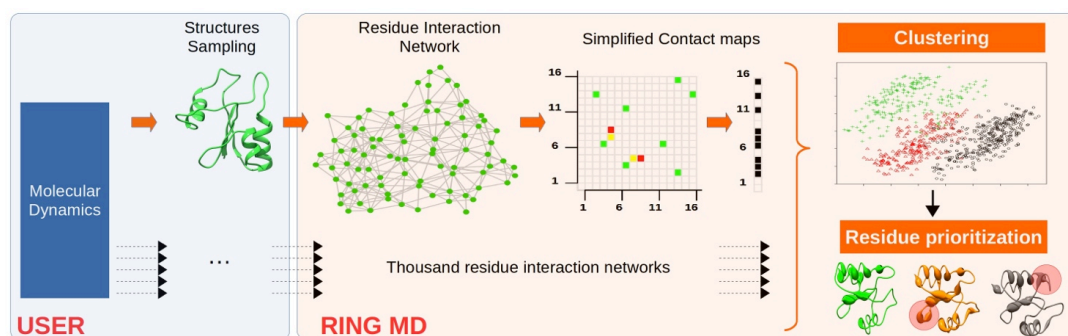

Figure 18: Starting from a MD, a user can easily sample PDB structures as representative of the whole simulation. With RING MD, the corresponding Residue Interaction Networks and contact maps are computed, and converted into a novel data structure called contact vector. Such representation emphasizes at the residue level the interacting amino acids by removing the sparse information about the contact partners. A clustering procedure then is used to select different conformers based on contact vectors. Structures belonging to highly diverse protein states are provided to the user, along with a prioritization of the amino acids explaining conformers diversity. Finally, the interaction frequency for each residue is reported in the selected structures, which summarize the intra-cluster contacts variability.

6.4 Results

A comparison between classical MD analysis and RING MD output is shown in Fig. 18. The RMSD [Fig. 18a] of two Ubiquitin (171) ionic interactions (residues K27-D52 and D39-R72) is compared with the RING MD summary structure [Fig. 19d-e]. We focused on R72 as it has widely been demonstrated to be involved in several ubiquitin interactions and cleavage phenomena (176–178). Particularly, in 2011 Ralat and co-workers demonstrated that Ubiquitin is a target of Insulin-degrading-enzyme (IDE). The catalytic activity of this enzyme passes through a first cleavage, involving the terminal glycine dimer and forming the 1-74 Ubiquitin segment, being R74 exposed to solvent. A second and slower cleavage follows the first one, and targets R72 forming the 1-72 Ubiquitin segment (176). In RING MD summary structure, residues are highlighted based on their interaction frequencies (see Fig. 19). The interaction between residues D39 and R72 is less frequent compared to K27-D52, which instead occurs almost in every frame. Therefore, according to RING MD, the slower IDE activity towards R72 could be explained by the formation for the 37% of the frames of R72-D39 ionic interaction, not allowing R72 to participate in the cleavage reaction. Indeed, IDE catalytic chamber has been proven to show protease activity towards highly flexible chains (179). Furthermore, RING MD provides an interaction-based trajectory clustering, with the conformers observed during the MD described by a linear combination of the contact-maps. In the reported case study, 4 clusters were obtained (see Fig. 19c). Particularly, structural clusters 3 and 4 represent molecular conformers not suitable for the cleavage reaction, containing R72 tightly locked with D39. On the contrary, in clusters 1 and 2 R72 is not involved in ionic interactions, and free to participate in other biochemical mechanisms. These protein states can clearly coexist, with the probability of observing them changing during time. Figure 19c represents the stochastic dynamic process, and facilitate in finding relationships among conformer transitions and periodicity. Extending the concept of RINs over a time dependent

scenario proved that the joint combination of multiple simplified contact maps contains all the initial RIN information. In fact, RING MD can reconstruct automatically the initial contacts with no information loss (see automatically coloured interacting residues in [Fig. 19b]). Moreover, the integration of the clustering results with the structural view can explain in detail the reasons of conformational changes. The output of the entire analysis is provided in supplementary information Fig. 1, highlighting changes in ionic, h-bond, and overall interactions. Conformational changes identified by the RING MD clustering show a good agreement with structure-based analysis (Sup Mat Tab 1) (176), and can be used to highlight the type of interactions relevant for these changes. It is important to note that classical clustering is based on 3D coordinates comparison of the structure, while TDCM compare graphs. Highly fluctuating residues, like N and C termini have high impact in terms of coordinates change, leading to a potential overestimation of structural distance. On the other hand, contact maps are robust to these irrelevant fluctuations and therefore more likely to provide a good final clustering. In this context, RING MD can play a key role in Conformational Ensembles for the analysis of intrinsic disorder proteins (180), since the classical structural comparison is known to be ineffective. We also observed that Inter-atomic interaction probabilities calculated with RING MD correlate significantly with Root Means Squared Fluctuations (RMSF) (Sup Mat Tab 2), thus explaining well the degrees of freedom of protein residues.
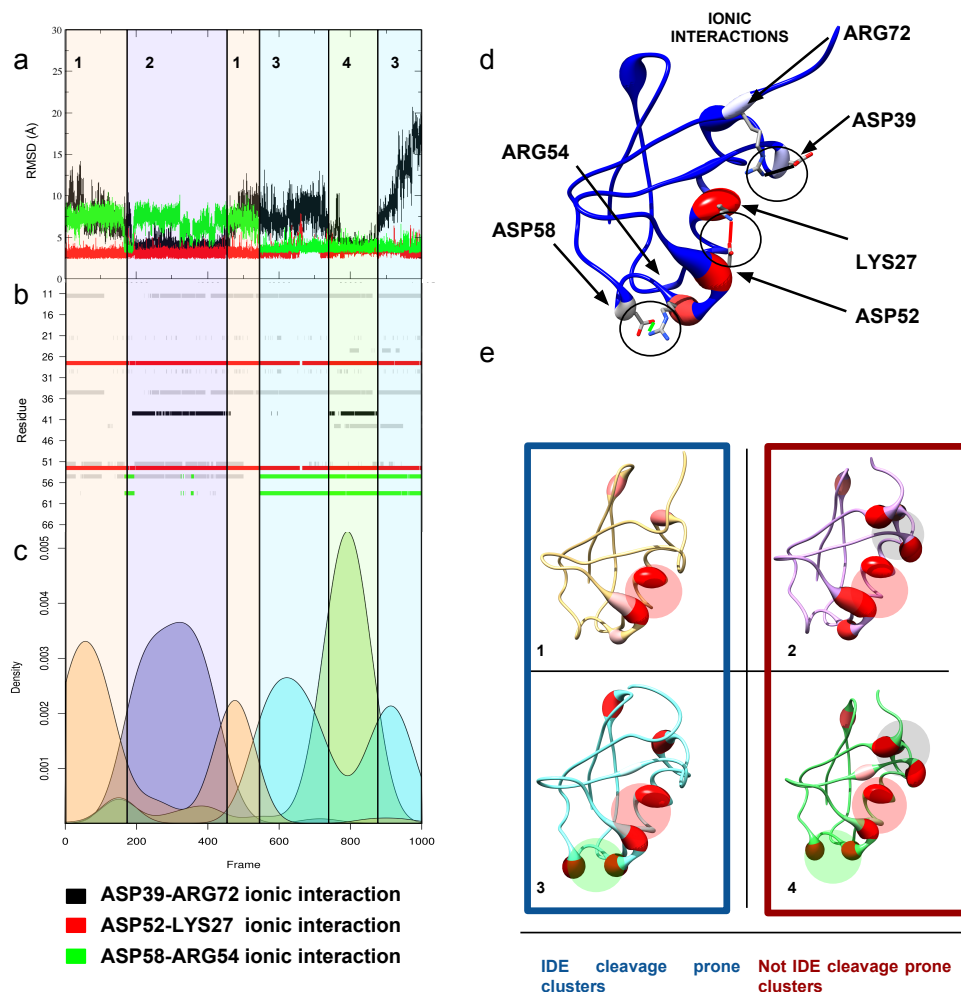
100

Figure 18: results of Ubiquitin RING MD analysis. a) Root mean square deviations (RMSDs) of ASP52-LYS27 (red) and ASP39-ARG72 (black), and ASP58-ARG54 (green) interactions. The distance in the plot is expressed in angstroms. b) Time Dependent Contact Map (TDCM) of the ionic interactions occurring in Ubiquitin MD simulation. Highlighted in red ASP52 and LYS27, in black ASP39-ARG72 and in green ASP58 and ARG54. c) Clusters deriving from k-means analysis of 1000 frames. a, b and c share the x axis, which is time (frames) of the MD simulation. The colouring scheme is the following: cluster 1 is light orange, cluster 2 is violet, cluster 3 is light blue and cluster 4 is light green. d) Summary structure reporting the frequencies of interaction between ASP39-ARG72 (light blue), LYS27-ASP52 (red) and ASP58-ARG54 (light blue and red, respectively). The thicker regions of this structure depict where ionic interactions (not highlighted) are occurring more frequently. e) Clusters structures. In cluster 1 ASP52-LYS27 is the only interaction occurring frequently. In cluster 2 ASP39-ARG72 and ASP52-LYS27 are occurring. In cluster 3 ASP52-LYS27 and ASP58-ARG54 are interacting and no interaction is found for ASP39 and ARG52. In cluster 4 all the chosen ionic couples are interacting.

6.5 Conclusion

The capability of molecular force-fields in simulating macromolecular systems provided a plethora of useful insights to scientific development. RING MD fits perfectly in this background, providing the user speed and reliability in analysing key features of a given system. The tool allows to quickly recognize structural hot spots in protein structures MD simulations, providing a fertile ground to improve protein biophysics knowledge in a faster and easier way. The results obtained with RING MD, compared to the literature, revealed that RING MD analysis could match hot spot residues with a simple trajectory analysis (42,178–184)(see Appendix).

7. Conclusions

During my Ph.D. experience I have been involved in several projects, ranging from pure biochemistry to biomedical sciences. In most cases, molecular dynamics simulations were at the basis of my works.

At first, with a great part of my work, highlights were proposed for VHL syndrome, starting from a new previously unreported role for hydroxyproline 567. The crystal structure of pVHL complex allowed the design of experiments to investigate on the role of this peculiar and still not well studied posttranslational modification. This clue gave a hint on the mechanism of what reported by Fedulova in 2007, suggesting a novel role for Prolyl-Hidroxylases (PHDs) enzymatic family. Since, by a multiple sequence alignment experiment of HIF-1α, we noticed that in several species of fish a serine residue is in the place of P567, we hypothesised a novel role in the oxygen-sensing pathway for this residue. Both hydroxyproline and serine contain indeed a hydroxyl moiety. The rearrangement of hydrogen bond networks within the pVHL-HIF-1α HYP567 chemical surrounding appeared plausible, and let us state that the pVHL residues involved in this new network could be considered as a putative novel binding surface. PHD-3 is expressed after a few hours of normoxyc conditions in cells where strong hypoxic stress occurred. This brought us to another question, still to be answered, about the involvement of PHD-3 in apoptosis. Molecular findings at this level shall be considered as suggestions, spots coming out from the messy cloud of cancer. The complex jigsaw of pVHL syndrome is relevant to a small percentage of the population, but still tracing the guidelines underneath this cancer is crucial. With these suggestions, in the framework of this thesis, we provided a small step forward towards the complete understanding of Von Hippel-Lindau syndrome driving mechanisms. The timescale we used to investigate this issue was nanoseconds. We performed 50 ns long simulations. In this time interval, as reported in Appendix 8.1, mutations of P567 with bulky residues brings to complex disruption. This implies another small, and in a certain

way obvious, answer. What we are used to consider a lifetime interval, which is supposed to be made of years, is driven by a huge amount of several different events occurring in timescales we are not even able to perceive. Enlarging the timescale, in the same subjects, we performed Petri net simulations of the, at the time of publishing (2014), almost entire pVHL known pathway. The transitions occurring in this second work, although the net was considered to be not time dependent, are belonging to a higher order of magnitude. In this case, what is noticeable is that apparently small changes, such as the lack of formation of a complex or a knockout of a single protein, could turn out to be fatal for cell life. The advantage of enlarging the view is that, by the usage of these particular networks, we were able to observe what could be thought as an interpretation of the Le Châtelier principle*. We were indeed able to observe all the transitions of the tokens caused by a system perturbation, undertaken by the system itself to minimize the damage brought by the change on the environment. In our case, the pVHL knockout system revealed that its role in oxygen sensing pathways, at least in short time ranges, is not sufficient to abolish HIF-1α degradation, as the latter shows to have back up reactions bringing to its degradation. In spite of this relatively damage brought to this particular pathway, pVHL deletion in the network showed to be in agreement with other secondary manifestations of the pVHL syndrome, unrelated to oxygen sensing. The enlargement of the network we built would, in case, provide a yet sounder framework to evaluate the impact of this molecular character.

Considering the impact that relatively small time intervals have on one complex and, in turn, on one pathway, an even smaller system can be brought as an example of small changes providing big effects. FAD is a relatively small molecule, compared with a massive protein structure such as Cryptochrome. This small molecule, weighing less than 800 Da, compared to dCry, weighing almost 60000 Da, have indeed a great as well as unexpected effect. The work we performed with this

---

* *Le Châtelier principle : Any change in a status quo prompts an opposing reaction in the responding system.*

complex highlighted how such small molecule can provide a necessary effect on a huge structure. Being the FAD completely buried within dCry structure, and being the overall volume of the FAD containing system and the system without FAD comparable, it carries out its structural role in chain flexibility. The concept of flexibility implies the concept of time, as the fluctuation is nothing but a difference between positions. What we observed with 20 ns long MD simulations was an unexpected increase in mean fluctuation rate for the complex with respect to the apoprotein. This can be seen as potential energy for the apoprotein transformed in kinetic energy for the complex. As the receptor activation requires the opening motion of one part of the protein, particularly the C-Terminal Tail, being in a more fluctuating situation could be intuitively thought as a strategy of nature to facilitate this event, and save time in reaching the transition state for activation. Light sensing is an event that shall not require long times to occur, otherwise becoming not useful. Due to the required computational power, we did not perform quanto-mechanical simulations of the light state of dCry. In fact, simulating FAD in radical semiquinone state could turn out very interesting and insightful, but shall require longer computational times to simulate shorter time ranges, in which we would probably see nothing occurring, at least simulating short nanoseconds long intervals.

In, talking about nanoseconds scale, short timescales what we can see is a structural relaxation of the target within the solvation box. In few nanoseconds, if a structure is not stable, the previously mentioned RMSD index tends to increase in a fast way. If one structure is, vice versa, stable, the RMSD index increases during the very beginning of the simulation, to later settle around a plateau value. In the case of the sequence permutations work, we predicted fifty different structures, trying to weigh the impact of different residue positions on sequences showing the same amino acidic percentage composition. The driving question was whether folds are kept because of evolution or they are the result of few stable folds formed randomly in the primordial broth and slightly improved by evolution. The shuffling

schema was used to simulate the heaviest mutations on different structures showing sequences made of the same percentage of residues. At the end of the analysis, it appeared that starting from the same amino acidic composition, different sequences show very different structures. To test this, we used *ab initio* structure predictions, and the obtained structures were all minimized through 4 ns long molecular dynamics simulations in water, in order to simulate an environment resembling as much as possible to the primordial broth. What is observable in nature is that the modern protein architectures are mostly combinations of a small number of protofolds. In this work, what arose is that using the same small building blocks, it is possible to obtain several different folds. The great advantage of this kind of investigations is that whenever we will collect enough information on the building blocks of proteins and their interconnections, we will probably be able to design *de novo* protein architectures to be exploited in several fields.

As one picture can not represent motion, so happens for crystal structures. Watching at a structure, which is not moving, can provide limited information. One structure shows indeed three dimensions. My last work, RING MD, focuses on this subject. The ratio under the tool is to provide information on motion by looking at a single representative snapshot. One molecule, being it a protein or a ligand or a DNA chain, whenever solvated moves around an equilibrium state, showing several different conformers. With RING MD, we performed contact based clusterization of three small targets trajectories. What we obtained were different conformations of the same proteins, characterized by different types of contacts. Two different conformers of one protein can act in different manners. A receptor can be in its active state, or inactive. A drug acts by locking a receptor in its active (agonist) or inactive (antagonist) state. Highlighting the structural key features of different pre-equilibrium states of the same protein obtained by a molecular dynamics trajectory clusterization can provide a fast insight on the behaviour of the system. What RING MD does is choosing a representative structure for every state resulting from the clusterization, and highlight the main interactions driving that peculiar conformer

by colour code. This provides a fast and immediate view of what happens during a molecular dynamics just with few structures. At the end, what we were able to do was trapping thousands of frames within few.

Finally, the roles that Time plays in biological system are several. Depending on the observed timescale, different phenomena can occur. The variable of Time is often taken for grant, as it is an integrating part of our lives that often we do not even pay attention to. Despite this, it is worth to notice, by the different experiments I have been involved in and that I did design myself, that Time not only plays a major role as "couch", but its role is active. It is not to be considered simply as something entraining the environment, it is what indeed modifies systems and environment. Different systems simply change in different ways, because of different mechanisms, but the main driving force should always be considered Time.

"We must use time as a tool, not as a couch"

J.F.Kennedy

8. Appendix of Supplementary Information

8.1 In silico investigation of PHD-3 specific HIF1-α proline 567 hydroxylation: a new player in the VHL/HIF-1α interaction pathway? Supplementary Information.

Root Mean Square Fluctuation (RMSF) analysis

In order to evaluate the effects of hydroxylation in P564 and P567 upon the complex stability, a RMSF plot for each protein in each run was carried out. The resulting graphs are shown in Supplementary Figure 5. The system containing no hydroxylation in either proline residues (Supplementary Figure 5-A) shows, in pVHL, a mean fluctuation around 0.25 Å, with peaks corresponding to regions between residues 69-83, 98-111, and 140-150, reaching values of 0.8 Å. The resulting value was considered standard and used to evaluate the remaining systems. In the same complex, HIF-1α shows a mean RMSF between 2 and 2.6 Å, which is in agreement with the importance of the previously reported hydroxylation in P564. The hydroxylation event in P564, according to the RMSF plots of the complex, appears to be stabilizing (Supplementary Figure 5-B). Indeed the mean RMSF of pVHL for this system fluctuates around 0.3Å, with peaks corresponding to the previously reported regions but reaching lower values. The result seems to confirm the functional role of this modification to reinforce the pVHL/HIF-1α interaction. HIF-1α segment RMSF is slightly higher than 2.1 Å until residues 501-503, than the value becomes lower than 2.0 Å. The data shows how the hydrogen bonds network surrounding HYP564 acts as a sort of lock, which retains the HIF-1α in closer contact than the unmodified proline residue. The hydroxylation event in P567 (Supplementary Figure 5-C) appears to be less stabilizing than the same event in P564, even if the RMSF plots of pVHL in the two complexes are comparable. Despite this, HIF-1α RMSF plot for this system shows

higher values, reaching peaks of 2.8 Å at the beginning of the crystallized segment, but lowering when getting closer to P567 probably due to the new-formed hydrogen bond network. The high RMSF value of the first residues is apparently related to the lack of usual hydrogen bonds network involving the hydroxylated P564. In the system containing both hydroxylated P564 and P567 (Supplementary Figure 5-D), pVHL manifestly gets more stable than all other analyzed systems, with a mean RMSF value lower than 0.25 Å, except for a peak corresponding to residues 160, 167-172 reaching values slightly higher than 0.5 Å, and a peak corresponding to residue 181 reaching a value of 0.75 Å. The corresponding double hydroxylated HIF-1α segment is much more constrained than in other systems, with a fluctuation between 0.8 and 0.2 Å. Analysis of the latter system shows a likely stabilizing role of hydroxyproline 567 in the complex, when compared to the other tested systems. Hydroxylation of only P567 appears to have no great role in the formation of the VHL-HIF-1 complex, but when occurring with P564 hydroxylation reinforcement of the complex stability seems to be relevant. Based on these results we tend to believe that hydroxylation in P567 should require the contemporary modification of P564 to be considered relevant for complex formation.

Control MD simulations

In order to check the pVHL/HIF-1α complex stability two control MD simulations were performed. The first one, derived from sequence analysis, containing the P567S mutant, the second one consisting in P564L and P567L mutants. RMSF calculations were performed for both systems.

*A*nalysis of the P567S system revealed a RMSF pattern for pVHL comparable with the plot of pVHL with hydroxyproline 564, showing plateaus between residues 70-80 slightly higher than 0.5 Å, between residues 100-111 reaching 0.6 Å, and 140-150, showing a peak RMSF value higher than 0.8 Å on residue 143 (Supplementary Figure 2-A). Interestingly, the RMSF of the HIF-1α segment showed a higher mean

value compared to the RMSF plot of HIF-1α containing hydroxyproline 564, floating around 3.0 Å and reaching peaks of almost 3.5 Å in residues 562-563. Thus, P567S cannot act as a hydroxyproline residue, but it suggests no destabilization of the pVHL/HIF-1α complex. In biological terms, we can assume this HIF-1α variant decrease hypoxia response in organisms living in a low oxygen environment.

The pVHL RMSF plot for the P564L–P567L double mutant system showed a pattern comparable with no hydroxyprolines (Supplementary Figure 2-B). Plateaus are present between residues 70-80, floating between 0.5 Å and 0.6 Å, between residues 100-110 floating around 0.6 Å, with a peak of 0.8 Å corresponding to residue 108. Other plateaus are noticeable between residues 140-160, with values around 0.5 Å and a peak of 0.8 Å corresponding to residue 144, and between residues 180-184, with average RMSF values around 0.5 Å. Interestingly, the corresponding HIF RMSF plot shows the highest values. The HIF-1α segment between residues 561-566 has RMSF values between 3.3 Å and 3.5 Å, and the remaining residues all have a RMSF value higher than 3.0 Å, showing the highest distance between the initial position of the HIF-1α segment and the pVHL of all tested systems. These two bulky residues substitutions highly destabilize the interaction between pVHL and HIF-1α (Supplementary Figure 3). We can speculate that in longer simulations the complex would probably dissociate. Finally, considering these results, we are sufficiently confident with the range of the simulation time scales used (i.e. nanoseconds).

```
RESI HYP              0.00
                                        HD1 HD2
ATOM N     N      -0.2900          |    \ /
ATOM CA    CP1     0.0200        N---CD  HG
ATOM HA    HB      0.0900          |     \  /
ATOM CD    CP3     0.1800          |      CG
ATOM HD1   HA      0.0000          |     / \
ATOM HD2   HA      0.0000       HA-CA--CB   OG2--HG2
ATOM C     C       0.5300          |   / \
ATOM O     O      -0.5300          | HB1 HB2
ATOM CB    CP2    -0.1800        O=C
ATOM HB1   HA      0.0900
ATOM HB2   HA      0.0900
ATOM CG    CP2     0.1400
ATOM HG    HA      0.0900
ATOM OG2   OH1    -0.6600
ATOM HG2   H       0.4300
```
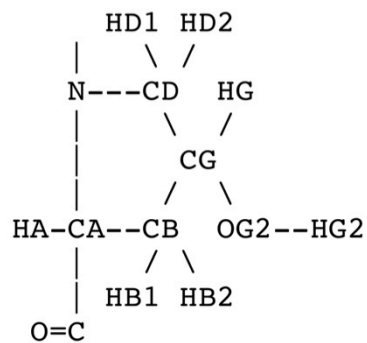
**Supplementary Figure 1. Atomic charges for HYP residues from the CHARMM force field.**

Atom names corresponding to the PRO residue were kept, while the hydroxyl oxygen was added. Atomic charges were derived from MOPAC package.
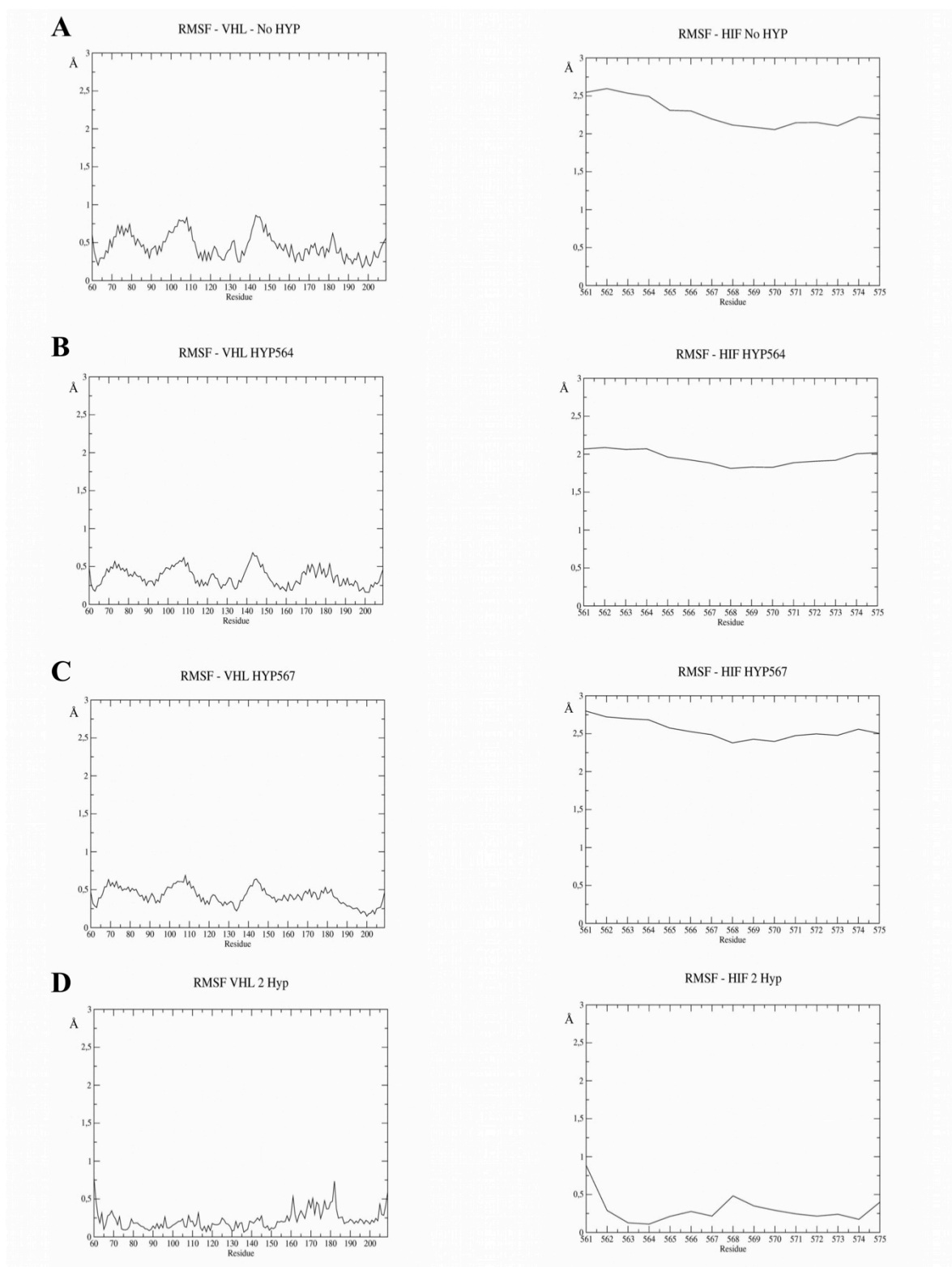
. **Supplementary Figure 2. RMSF plots of the tested complexes.** Figures show the root mean square fluctuation (RMSF, y axis) against residue number (x axis) of the complexes (A) with no hydroxylation on P564 and P567, (B) hydroxylated P564, (C) hydroxylated P567, and (D) both P564 and P567 hydroxylated

114
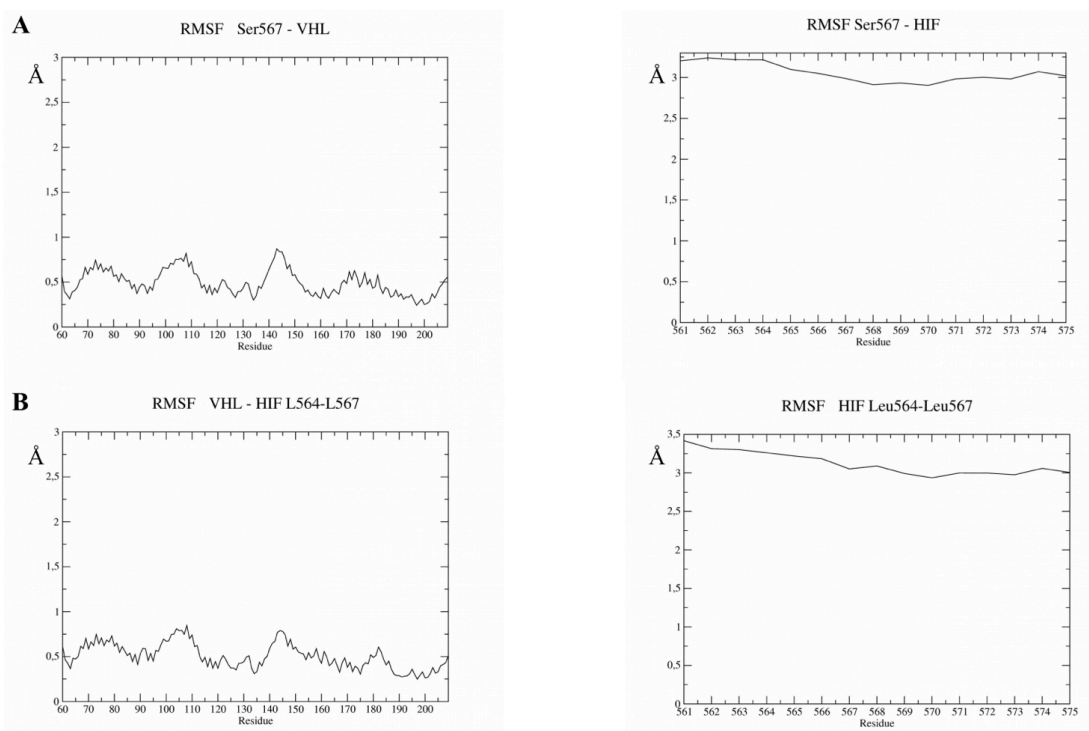
Supplementary Figure 3. RMSF plots of the control MD simulations. Figures show the root mean square fluctuation (RMSF, y axis) against residue number (x axis) of the P567S mutant and the P564L-P567L double mutant systems.
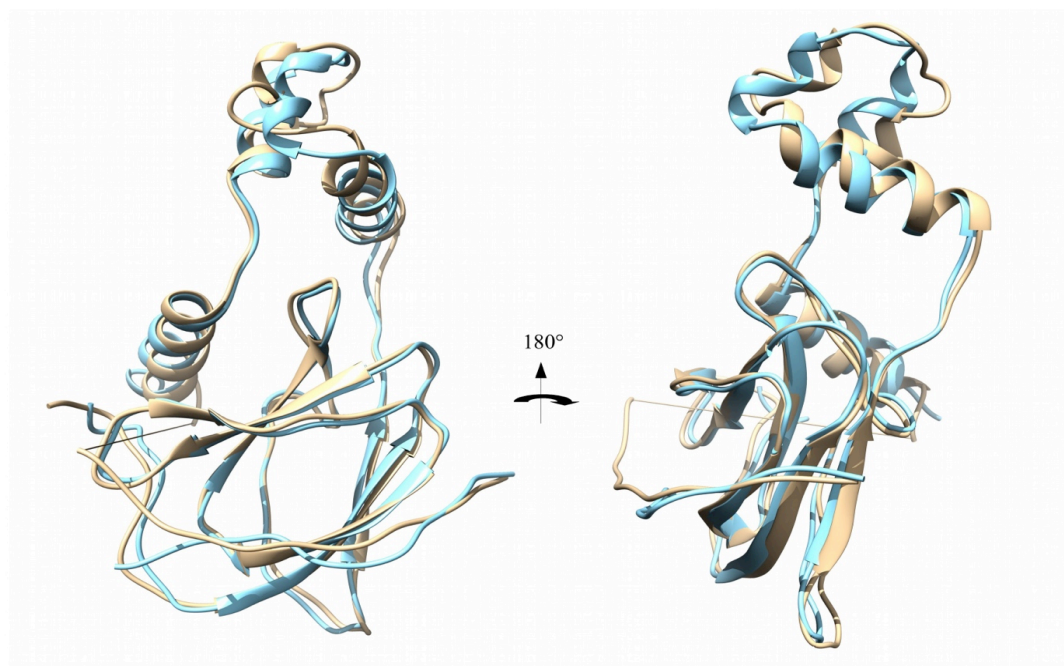
Supplementary Figure 4. Wild-type pVHL/HIF-1α complex vs. P564L-P567L double mutant after the MD simulation. The figure shows a superimposition of the crystal structure of the pVHL/HIF-1α complex, (blue, PDB code 1LM8), and the P564L-P567L mutant complex after 50 ns of MD simulation (light brown).

8.2 Design and analysis of a Petri net model of the Von Hippel-Lindau (VHL) tumor suppressor interaction network. Supplementary Information.

ID	Name	Pre-Places and Post-Places	Biological Meaning
t_0	_trans_0	[Hif_IN + 1] & [hif - 1];	Entrance of HIF in nucleus
t_1	_trans_1	[_place_2 + 1] & [Hif_IN - 1] & [arnt - 1];	Dimerization of HIF
t_2	_trans_2	[hif + 1] & [Hif_IN - 1];	Export of HIF from nucleus
t_3	_trans_3	[_place_6 + 1] & [_place_2 - 1] & [p300 - 1] & [creb - 1];	Complex formation of HIF and other coactivators of transcriprion
t_4	_trans_4	[_place_7 + 1] & [_place_6 - 1] & [cjun1 - 1];	Complex with cjun coactivator
t_5	_trans_5	[_place_9 + 1] & [_place_7 - 1] & [hre - 1];	Attach to DNA in HRE segment

		[_place_10 + 1] & [hre + 1] & [_place_11 + 1] & [_place_12 + 1] & [_place_13 + 1] & [_place_14 + 1] & [cjun1 + 1] & [p300 + 1] & [arnt + 1] & [creb + 1] & [Hif_IN + 1] & [_place_218 + 1] &	
t_6	_trans_6	[_place_9 - 1];	Transcription of growth factors' mRNA and others
t_7	_trans_7	[VEGF + 1] & [_place_11 - 1];	Formation of protein from mRNA
t_8	_trans_8	[EPO + 1] & [_place_12 - 1];	Formation of protein from mRNA
t_9	_trans_9	[ET1 + 1] & [_place_10 - 1];	Formation of protein from mRNA
t_10	_trans_10	[LDH + 1] & [_place_13 - 1];	Formation of protein from mRNA
t_11	_trans_11	[pyrDHK + 1] & [_place_14 - 1];	Formation of protein from mRNA
t_12	_trans_12	[rec_act + 1] & [_place_19 - 1] & [VEGF - 1];	Activation of GF receptor
t_13	_trans_13	[_place_19 + 1] & [_place_21 + 1] & [rec_act - 1] & [PLC1 - 1];	Activation of PLC by VEGF receptor
t_14	_trans_14	[_place_19 + 1] & [PI3k_act + 1] & [rec_act - 1] & [PI3K - 1];	Activation of PI3K by VEGF receptor
t_15	_trans_15	[_place_19 + 1] &	Activation of Ras

		[_place_24 + 1] & [_place_55 + 1] & [ADP + 1] & [rec_act - 1] & [RAS1 - 1] & [gtp1 - 1] & [ATP - 1];	GTP-ase by VEGF receptor
t_16	_trans_16	[ip31 + 1] & [DAG1 + 1] & [PLC1 + 1] & [PI2P + 1] & [PI2P - 1] & [_place_21 - 1];	Breaking of PI2P in DAG and IP3 by PLC
t_17	_trans_17	[Ca_chan1 + 1] & [ca1 + 1] & [ip31 - 1] & [Ca_chan1 - 1];	Opening of Calcium channel by IP3
t_18	_trans_18	[_place_32 + 1] & [cam_k_kk1 + 1] & [ca1 - 1] & [cam_k_kk1 - 1];	Formation of complex calmodulin kinase
t_19	_trans_19	[_place_31 + 1] & [PKC1 - 1] & [DAG1 - 1];	Activation of PKC by DAG
t_20	_trans_20	[PKC1 + 1] & [_place_33 + 1] & [_place_31 - 1];	Signal tranduction by PKC
t_21	_trans_21	[PI3K + 1] & [PIP3 + 1] & [PI2P + 1] & [PI3k_act - 1] & [PI2P - 1];	PI3K phosphorilates PI2P in PI3P
t_22	_trans_22	[RAS1 + 1] & [_place_49 + 1] & [_place_24 - 1] & [mapkkk - 1];	Ras starts MAPK cascade

t_23	_trans_23	[pdk_PP + 1] & [PKB_PP + 1] & [PIP3 - 1] & [PDK - 1] & [PKB - 1];	PI3P activates PKB and PDK via phosphorilation
t_24	_trans_24	[PDK + 1] & [pkb_act + 1] & [pdk_PP - 1] & [PKB_PP - 1];	PDK activates completely PKB to act on apoptosis signal
t_25	_trans_25	[BAP_DIP + 1] & [inhib + 1] & [_place_43 - 1] & [bad_inact - 1];	BAD-DIP complex formation
t_26	_trans_26	[bad_inact + 1] & [inhib - 1] & [BAD - 1];	Inactivation of BAD via inhibitor
t_27	_trans_27	[_place_43 + 1] & [_place_47 + 1] & [DIP - 1];	Signal of survival by DIP
t_28	_trans_28	[mapkkk + 1] & [_place_51 + 1] & [_place_49 - 1] & [mapkk - 1];	MAPK cascade, activating MAPKK
t_29	_trans_29	[mapkk + 1] & [_place_53 + 1] & [_place_51 - 1] & [mapk - 1];	MAPK cascade, activating MAPK
t_30	_trans_30	[mapk + 1] & [_place_59 + 1] & [_place_53 - 1];	Proliferation signal by MAPK cascade
t_31	_trans_31	[DIP + 1] & [BAD + 1] & [PKB + 1] & [pkb_act - 1] & [BAP_DIP - 1];	PKB breaks BAD DIP binding to avoid apoptosis

t_32	_trans_32	[EPOR + 1] & [_place_61 - 1] & [EPO - 1];	Activation of EPO receptor by Erythropoietin
t_33	_trans_33	[PLC_ACT + 1] & [EPOR - 1] & [_place_63 - 1];	Activation of PLC by EPOR
t_34	_trans_34	[STATx2 + 1] & [_place_63 + 1] & [_place_61 + 1] & [PLC_ACT - 1] & [stat5 - 1];	PLC allows formation of dimer STAT5 activating monomer
t_35	_trans_35	[jak2 + 1] & [_place_61 + 1] & [EPOR - 1] & [jak - 1];	Activation of jak by EPOR
t_36	_trans_36	[STATx2 + 1] & [jak + 1] & [jak2 - 1] & [stat5 - 1];	Jak allows formation of dimer STAT5 activating monomer
t_37	_trans_37	[stat5 + 2] & [_place_89 + 1] & [_place_87 + 1] & [STATx2 - 2] & [_place_87 - 1];	STAT5 dimer transcription activity
t_38	_trans_38	[_place_70 + 1] & [_place_61 + 1] & [EPOR - 1] & [shc2 - 1];	Activation of shc by EPOR
t_39	_trans_39	[complex1 + 1] & [_place_70 - 1] & [SOS12 - 1] & [GRB2 - 1];	Complex formation of sos1 and grb2

t_40	_trans_40	[SOS12 + 1] & [GRB2 + 1] & [RAS3 + 1] & [_place_83 + 1] & [shc2 + 1] & [complex1 - 1] & [GTP2 - 1] & [_place_82 - 1];	The SOS and GRB complex activates ras protein
t_41	_trans_41	[GTP2 + 1] & [RAF2 + 1] & [_place_82 + 1] & [ADP + 1] & [RAS3 - 1] & [_place_76 - 1] & [_place_83 - 1] & [ATP - 1];	Ras activates raf protein with consumption of ATP
t_42	_trans_42	[_place_76 + 1] & [MEk1 + 1] & [RAF2 - 1] & [_place_78 - 1];	MEK/ERK/ELK cascade
t_43	_trans_43	[_place_78 + 1] & [ERK1 + 1] & [MEk1 - 1] & [_place_80 - 1];	Erk is activated by MEk
t_44	_trans_44	[_place_84 + 1] & [ERK1 - 1];	Erk activate ELK
t_45	_trans_45	[ELK1 + 1] & [_place_80 + 1] & [_place_84 - 1] & [_place_86 - 1];	ELK is moved in the nucleus while separating from ERK
t_46	_trans_46	[_place_88 + 1] & [_place_86 + 1] & [_place_87 + 1] & [ELK1 - 1] & [_place_87 - 1];	Transcription activity for Cfos
t_47	_trans_47	[CFOS + 1] & [_place_88 - 1];	mRNA translation for CFOS protein

122

		[_place_92 + 1] & [CFOS - 1] & [cjun - 1];	
t_48	_trans_48		Complex of cfos and cjun
t_49	_trans_49	[cjun + 1] & [_place_95 + 1] & [_place_93 - 1] & [_place_95 - 1];	Activation of cjun
t_50	_trans_50	[_place_93 + 1] & [_place_94 + 1] & [_place_87 + 1] & [_place_87 - 1] & [_place_92 - 1];	Transcription activity cjun/cfos
t_51	_trans_51	[_place_98 + 1] & [ET1 - 1] & [ETRa2 - 1];	Activation of ET receptor b by endotelin
t_52	_trans_52	[_place_97 + 1] & [ET1 - 1] & [ETRa - 1];	Activation of ET receptor a
t_53	_trans_53	[ETRa2 + 1] & [_place_100 + 1] & [_place_97 - 1] & [PG - 1];	Activation of PG by ETRa
t_54	_trans_54	[ETRb + 1] & [_place_101 + 1] & [PI3k + 1] & [csrc1 + 1] & [_place_98 - 1] & [shc1 - 1] & [PI3k - 1] & [csrc1 - 1];	ETRb, PI3K and csrc activates shc1
t_55	_trans_55	[PG + 1] & [_place_105 + 1] & [_place_100 - 1] & [PLC2 - 1];	PG activate PLC

t_56	_trans_56	[_place_110 + 1] & [DAG - 1] & [PKC2 - 1];	DAG activates PKC
t_57	_trans_57	[PKC2 + 1] & [_place_114 + 1] & [RAs2 + 1] & [ADP + 1] & [_place_110 - 1] & [ATP - 1] & [raf2 - 1] & [_place_111 - 1];	PKC activates raf together with ras to start MAPK cascade
t_58	_trans_58	[raf2 + 1] & [_place_116 + 1] & [_place_114 - 1] & [mapkk2 - 1];	Raf activates MAPKK
t_59	_trans_59	[mapkk2 + 1] & [_place_118 + 1] & [_place_116 - 1] & [mapk2 - 1];	MAPKK activates MAPK
t_60	_trans_60	[mapk2 + 1] & [Et_eff2 + 1] & [_place_313 + 1] & [_place_118 - 1] & [_place_313 - 1];	MAPK generates effects of vasocostriction and high pressure
t_61	_trans_61	[DAG + 1] & [PLC2 + 1] & [IP3 + 1] & [pip + 1] & [pip - 1] & [_place_105 - 1];	PLC creates DAG and IP3 from PIP
t_62	_trans_62	[camk_pp + 1] & [ca - 1] & [cam_k_kk2 - 1];	Calcium allows binding of calmodulin and kinases
t_63	_trans_63	[CREB_P + 1] & [cam_k_kk2 + 1] & [_place_123 - 1] & [camk_pp - 1];	Calmodulin kinase activates CREB, transcription factor

124

t_64	_trans_64	[_place_123 + 1] & [Et_eff1 + 1] & [_place_313 + 1] & [CREB_P - 1] & [_place_313 - 1];	CREB generates other effects of endothelin
t_65	_trans_65	[complex + 1] & [sos1 - 1] & [grb1 - 1] & [_place_101 - 1];	Complex formation of shc, sos and GRB2
t_66	_trans_66	[shc1 + 1] & [grb1 + 1] & [sos1 + 1] & [_place_111 + 1] & [complex - 1] & [RAs2 - 1];	Activation of ras by grb/sos/shc complex
t_67	_trans_67	[ca + 1] & [_place_314 + 1] & [IP3 - 1] & [_place_314 - 1];	IP3 opens calcium channels from deposits
t_68	absorbed	[_place_130 + 1] & [glu1 - 1];	Glucose enters cells for glycolysis
t_69	eating	[glu1 + 1];	Glucose assimilation from food
t_70	_trans_70	[G6P + 1] & [ADP + 1] & [exokinase + 1] & [_place_130 - 1] & [ATP - 1] & [exokinase - 1];	Glucose Phosphorylation
t_71	_trans_71	[F6P + 1] & [Isomerase + 1] & [G6P - 1] & [Isomerase - 1];	G6P is modified in Fructose-6-phosphate

t_72	_trans_72	[F16BP + 1] & [PFK + 1] & [ADP + 1] & [F6P - 1] & [PFK - 1] & [ATP - 1];	F6P is phosphorylated
t_73	_trans_73	[DHAP + 1] & [ALDO + 1] & [GADP + 1] & [F16BP - 1] & [ALDO - 1];	F16BP is broken in Glyrealdehyde 3-Phosphate and dihydroxyacetone phosphate
t_74	_trans_74	[TPI + 1] & [GADP + 1] & [DHAP - 1] & [TPI - 1];	DAHP is modified in GADP
t_75	_trans_75	[GAPDH + 1] & [BPG + 1] & [NADH + 1] & [P + 1] & [GADP - 1] & [GAPDH - 1] & [NAD - 1] & [P - 1];	GADP dehydrogenation to 1,3-biphosphoglycerate
t_76	_trans_76	[ATP + 1] & [PG3 + 1] & [PGK + 1] & [BPG - 1] & [ADP - 1] & [PGK - 1];	Dephosphorylation in 3-phosphoglycerate
t_77	_trans_77	[ENO + 1] & [pep + 1] & [ENO - 1] & [PG2 - 1];	Formation of Phosphoenolpyruvate
t_78	_trans_78	[ATP + 1] & [Pyr_kin + 1] & [pyr + 1] & [pep - 1] & [Pyr_kin - 1] & [ADP - 1];	Hydrolization of PEP in pyruvate, formation of a ATP
t_79	_trans_79	[lactic_acid + 1] & [pyr - 1] & [LDH - 1];	Pyruvate to lactic acid
t_80	_trans_80	[_place_152 + 1] & [oxa + 1] & [pyr - 1] & [_place_152 - 1];	Pyruvate to oxalacetate for carboxylation

t_81	_trans_81	[COA + 1] & [AChCOA + 1] & [_place_157 + 1] & [pyr - 1] & [COA - 1] & [_place_156 - 1];	Pyr to Acetyl-coenzyme A
t_82	_trans_82	[_place_156 + 1] & [_place_157 - 1];	Deactivation of Pyr-dehydrogenase
t_83	_trans_83	[_place_157 + 1] & [pyrDHK - 1] & [_place_156 - 1];	PyrDH activation by Kinase
t_84	_trans_84	[cit_synt + 1] & [Cyt + 1] & [oxa - 1] & [AChCOA - 1] & [cit_synt - 1];	Krebs step 1: citrate
t_85	_trans_85	[cisAco + 1] & [Cyt - 1] & [aconitase - 1];	Krebs step 2: cis-aconitate
t_86	_trans_86	[isocyt + 1] & [aconitase + 1] & [cisAco - 1];	Krebs step 3: isocitrate
t_87	_trans_87	[oxasu + 1] & [NADH + 1] & [isocyt - 1] & [isocyt_DH - 1] & [NAD - 1];	Krebs step 4a: oxalsuccinate
t_88	_trans_88	[a_cheto + 1] & [isocyt_DH + 1] & [oxasu - 1];	Krebs step 4b: a-chetoglutarate
t_89	_trans_89	[SuCoA + 1] & [acheto_DH + 1] & [NADH + 1] & [a_cheto - 1] & [acheto_DH - 1] & [NAD - 1];	Krebs step 5: SuccinilCoA

t_90	_trans_90	[su + 1] & [SuCoA_synth + 1] & [GTP3 + 1] & [SuCoA - 1] & [SuCoA_synth - 1] & [p2 - 1] & [GDP2 - 1];	Krebs step 6: Succinate
t_91	_trans_91	[fum + 1] & [su_DH + 1] & [su - 1] & [su_DH - 1];	Krebs step 7: fumarate
t_92	_trans_92	[mal + 1] & [fumarase + 1] & [fum - 1] & [fumarase - 1];	Krebs step 8 malate
t_93	_trans_93	[oxa + 1] & [mal_DH + 1] & [mal - 1] & [mal_DH - 1];	Krebs step 9 oxalacetate
t_94	_trans_94	[glut2 + 1] & [_place_176 + 1] & [glu1 - 1] & [glut2 - 1];	Activation of GLUT2
t_95	_trans_95	[insulin + 1] & [_place_188 + 1] & [_place_176 - 1] & [_place_187 - 1];	Synthesis of insulin
t_96	_trans_96	[glicogen + 1] & [GS_act + 1] & [GS_act - 1] & [glu2 - 1];	Formation of glycogen
t_97	_trans_97	[H2O + 1] & [ATP + 1] & [NAD + 1] & [NADH - 1] & [O2 - 1] & [ADP - 1];	ATP synthesis
t_98	_trans_98	[vhl + 1] & [hif + 1];	Input for VHL and HIF

t_99	_trans_99	[GS_act + 1] & [_place_180 + 1] & [insulin + 1] & [GS_inact - 1] & [_place_180 - 1] & [insulin - 1];	Activation of Glycogen Synthetase by insulin
t_100	_trans_100	[GS_inact + 1] & [gsk3b_act + 1] & [GS_act - 1] & [gsk3b_act - 1];	Inactivation of GS by GSK3β
t_101	_trans_101	[PKB_akt + 1] & [PKB_AKT - 1];	Protein Kinase B inactivation
t_102	_trans_102	[PKB_AKT + 1] & [insulin + 1] & [PKB_akt - 1] & [insulin - 1];	Activation of PKB by insulin
t_103	_trans_103	[gsk3b_act + 1] & [P_ase + 1] & [gsk3b_inact - 1] & [P_ase - 1];	Activation of GSK3β by phosphatase
t_104	_trans_104	[_place_187 + 1] & [_place_188 - 1];	Inhibition of insulin production
t_105	glut4_musc	[insulin + 1] & [glu2 + 1] & [insulin - 1] & [_place_176 - 1];	Insulin activation of GLUT4
t_106	_trans_106	[gsk3b_act + 1] & [hif_ppp + 1] & [hif - 1] & [gsk3b_act - 1];	Phosphorylation of Hif by GSK3β
t_107	_trans_107	[ub_ase + 1] & [hif_ppp_ub + 1] & [ub_ase - 1] & [hif_ppp - 1];	Ubiquitination of phosphorylated Hif
t_108	degradation2	[hif_ppp_ub - 1];	Degradation of hif via

			GSK
t_109	_trans_109	[gsk3b_act + 1] & [vhl_pp + 1] & [gsk3b_act - 1] & [vhl - 1];	Phosphorylation of Vhl by GSK
t_110	mt_instab	[tubulin + 1] & [vhl_pp - 1] & [mt - 1];	Destabilization of microtubules
t_111	_trans_111	[PHD2_act + 1] & [O2 - 1] & [_place_195 - 1];	Activation of PHD2 by oxygen
t_112	_trans_112	[_place_195 + 1] & [PHD2 - 1] & [_place_205 - 1];	Inhibition of PHD by Krebs species
t_113	_trans_113	[_place_197 + 1] & [O2 - 1] & [ARD - 1];	Activation of ARD by oxygen
t_114	_trans_114	[PHD2_act + 1] & [O2 - 1] & [PHD2 - 1];	Activation of PHD2 by oxygen
t_115	_trans_115	[ARD + 1] & [PHD2 + 1] & [Hy_Ac_HIF1a + 1] & [hif - 1] & [PHD2_act - 1] & [_place_197 - 1];	Modification of Hif in normoxia
t_116	_trans_116	[vcb + 1] & [elob + 1] & [eloc + 1] & [elob - 1] & [eloc - 1] & [vhl - 1];	Formation of Vcb complex
t_117	_trans_117	[vcd + 1] & [Cu2 + 1] & [vcb - 1] & [Cu2 - 1];	Interaction with Cu2, formation of Vcd
t_118	_trans_118	[Hif_ub + 1] & [vcd - 1] & [Hy_Ac_HIF1a - 1];	Ubiquitination of modified Hif
t_119	degradation1	[Hif_ub - 1];	Degradation of Hif

t_120	_trans_120	[_place_205 + 1] & [oxa - 1];	Uptake of oxalacetate for PHD2 inhibition
t_121	_trans_121	[_place_205 + 1] & [fum - 1];	Uptake of malate for PHD2 inhibition
t_122	_trans_122	[_place_205 + 1] & [su - 1];	Uptake of succinate for PHD2 inhibition
t_123	_trans_123	[O2 + 2] & [Et_eff1 - 1];	Creation of oxygen by Endothelin effect
t_124	_trans_124	[O2 + 2] & [_place_33 - 1];	Creation of oxygen by VEGF
t_125	_trans_125	[O2 + 2] & [_place_59 - 1];	Creation of oxygen by Vegf
t_126	_trans_126	[O2 + 2] & [_place_47 - 1];	Creation of oxygen by Vegf
t_127	_trans_127	[O2 + 2] & [_place_94 - 1];	Creation of oxygen by EPO
t_128	_trans_128	[O2 + 2] & [_place_32 - 1];	Creation of oxygen by Vegf
t_129	_trans_129	[O2 + 2] & [Et_eff2 - 1];	Creation of oxygen by Endothelin
t_130	_trans_130	[O2 + 2] & [_place_89 - 1];	Creation of oxygen by EPO
t_131	_trans_131	[p2 + 1] & [GDP2 + 1] & [GTP3 - 1];	Breaking of GTP
t_132	_trans_132	[gsk3b_inact + 1] & [PKB_AKT + 1] & [PKB_AKT - 1] & [gsk3b_act - 1];	Inactivation of GSK by PKB
t_133	_trans_133	[mt + 1] & [tubulin - 1] & [vhl - 1];	Stabilization of microtubules
t_134	_trans_134	[nedd_vhl + 1] & [nedd8 - 1] & [vcb - 1];	Neddylation of Vhl

t_135	_trans_135	[VHL_FN + 1] & [FN - 1] & [nedd_vhl - 1];	Interaction of Vhl and Fibronectin
t_136	matrix	[stable_matrix + 1] & [VHL_FN - 1];	Stabilization of ECM
t_137	_trans_137	[FN + 1] & [nedd8 + 1] & [vcb + 1] & [stable_matrix - 1] & [MT1MMP - 1];	Destabilization of ECM by MMP
t_138	_trans_138	[MT1MMP + 1] & [_place_218 - 1];	Production of MMP
t_139	_trans_139	[mmp_inact + 1] & [_place_221 + 1] & [MT1MMP - 1] & [timp1 - 1];	Inhibition of MMP by TIMP
t_140	_trans_140	[timp1 + 1] & [_place_221 - 1] & [vcb - 1];	Activation of TIMP by Vhl
t_141	_trans_141	[insulin - 1];	Elimination of insulin
t_142	_trans_142	[gtp1 + 1] & [_place_55 - 1];	Formation of GTP
t_143	_trans_143	[prot_no_ub + 1] & [prot_ub - 1] & [VDU - 1];	Deubiquitination by VDU
t_144	VDU_degr	[_place_227 - 1];	VDU degradation by VHL
t_145	_trans_145	[_place_227 + 1] & [Cu2 + 1] & [Cu2 - 1] & [_place_225 - 1];	Binding of vcb and VDU with Cu2
t_146	_trans_146	[_place_225 + 1] & [vcb - 1] & [VDU - 1];	Interaction of VDU and VHL
t_147	out_nucleus	[_place_228 - 1];	Extrusion of VHL by eEF1A

t_148	_trans_148	[_place_228 + 1] & [_place_231 + 1] & [eEF1a + 1] & [rpb1 + 1] & [rpb7 + 1] & [eEF1a - 1] & [vhl - 1] & [_place_230 - 1];	Activation of eEF1A by RNA pol II to extrude Vhl
t_149	_trans_149	[_place_230 + 1] & [_place_231 - 1] & [rpb7 - 1] & [rpb1 - 1];	Formation of RNA pol II from subunits
t_150	_trans_150	[_place_235 + 1] & [card9 - 1] & [nfkb - 1];	Binding of Card9 with NFkB
t_151	_trans_151	[tnfa + 1] & [nfkb + 1] & [card9 + 1] & [_place_236 + 1] & [_place_235 - 1] & [_place_236 - 1];	Stimulation of TNFα production
t_152	_trans_152	[_place_238 + 1] & [card9 - 1] & [vcb - 1];	Binding Vhl and Card9
t_153	_trans_153	[_place_240 + 1] & [_place_238 - 1] & [ck2 - 1];	Phosphorylation of card9 (A)
t_154	_trans_154	[card9_p + 1] & [ck2 + 1] & [_place_240 - 1];	Phosphorylation of card9 (B)
t_155	_trans_155	[PHD2_act + 1] & [phy_rpb1 + 1] & [_place_242 - 1] & [PHD2_act - 1];	Hydroxylation of RPB1 by PDH2
t_156	_trans_156	[Rpb_hy_p + 1] & [rpb_kin + 1] & [phy_rpb1 - 1] & [rpb_kin - 1];	Phosphorylation of Hydroxylated RPB1

		[ub_moved + 1] & [ub + 1] & [ub - 1] & [Rpb_hy_p - 1];	Movement of RPB1 to other area of nucleus
t_157	_trans_157		
t_158	_trans_158	[_place_242 + 1] & [rpb1 - 1] & [vcb - 1];	Binding of Vhl to RPB1
t_159	_trans_159	[Rpb7_vhl + 1] & [rpb7 - 1] & [vcb - 1];	Binding of Vhl to RPB7
t_160	Rpb_degr	[Rpb7_vhl - 1];	Degradation of RPB7
t_161	_trans_161	[stab_complex + 1] & [jade1 - 1] & [vhl - 1];	Stabilization of Jade1 by Vhl
t_162	transcription	[proliferation + 1] & [_place_309 - 1];	Wnt proliferative signal
t_163	_trans_163	[b_cat_P + 1] & [_place_258 + 1] & [gsk3b_act + 1] & [_place_257 + 1] & [GSK3B_complex - 1] & [b_catenin - 1];	Phosphorylation of b-catenin and decomposition of kinase-complex
t_164	_trans_164	[ub_jade + 1] & [b_cat_P - 1] & [stab_complex - 1];	Jade bind b-catenin for degradation
t_165	_trans_165	[_place_309 + 1] & [wnt + 1] & [wnt - 1] & [b_catenin - 1];	Activtion of b-catenin
t_166	_trans_166	[GSK3B_complex + 1] & [gsk3b_act - 1] & [_place_257 - 1] & [_place_258 - 1];	Formation of kinase complex from axin, APC and GSK
t_167	_trans_167	[pkcz2 + 1] & [sp1_P + 1] & [pkcz2 - 1] & [sp1_solo - 1];	Phosphorylation of Sp1 by PKCζ
t_168	PKCz2_degr	[_place_318 - 1];	Degradation of PKCζ

t_169	_trans_169	[_place_259 + 1] & [pkcz2 - 1] & [vcb - 1];	Binding of Vhl and PKCζ
t_170	_trans_170	[_place_263 + 1] & [par6 - 1] & [aPKC - 1];	Activation of aPKC by Par6
t_171	_trans_171	[aPKC_act + 1] & [_place_263 - 1] & [cdc42 - 1];	Complete activation of aPKC via cdc42
t_172	_trans_172	[par3_P + 1] & [cdc42 + 1] & [aPKC + 1] & [aPKC_act - 1] & [par3 - 1];	Formation of par complex
t_173	_trans_173	[tight_j_form + 1] & [par3 + 1] & [par6 + 1] & [par3_P - 1];	Formation of tight junctions
t_174	_trans_174	[par6_pkcz + 1] & [pkcz2 - 1] & [par6 - 1];	Binding par6 and PKCζ
t_175	_trans_175	[nur_vhl_hif + 1] & [Hif_IN - 1] & [nur_vhl - 1];	Formation of Nur77-Vhl-Hif complex
t_176	_trans_176	[nur77 + 1] & [nur77_mRNA - 1];	Translation of Nur77 mRNA
t_177	_trans_177	[sp1_vhl + 1] & [vhl - 1] & [sp1_solo - 1];	Binding Sp1 and Vhl
t_178	_trans_178	[vegf_stab + 1] & [vegf2 - 1] & [hur - 1];	Stabilization of Vegf by HuR
t_179	transcription_2	[sp1_P + 1] & [vegf2 + 1] & [sp1_P - 1];	Transcription of Vegf by Sp1
t_180	_trans_180	[nur_vhl + 1] & [nur77 - 1] & [vhl - 1];	Binding Nur77 and Vhl

t_181	_trans_181	[no_stab + 1] & [hur - 1] & [vhl - 1];	Inhibition of stabilization action by Hur via Vhl
t_182	_trans_182	[stabilization + 1] & [hur - 1] & [_place_315 - 1];	Stabilization of IGFR mRNA by Hur
t_183	_trans_183	[IGF1R + 1] & [hur + 1] & [stabilization - 1];	Translation for IGF1R
t_184	_trans_184	[igfr_act + 1] & [IGF1 + 1] & [IGF1 - 1] & [IGF1R - 1];	Activation of receptor by IGF1
t_185	_trans_185	[pi3k_act + 1] & [IGF1R + 1] & [pi3k - 1] & [igfr_act - 1];	Activation of PI3k by IGFR
t_186	_trans_186	[pip3 + 1] & [pi3k + 1] & [pi3k_act - 1];	Formation of PIP3
t_187	_trans_187	[AKT_mem + 1] & [pip3 - 1] & [akt - 1];	Movement of AKT to membrane
t_188	_trans_188	[AKt_PP + 1] & [PDPK1_2 + 1] & [mtorc2_1 + 1] & [mtorc2_1 - 1] & [AKT_mem - 1] & [PDPK1_2 - 1];	Phosphorylation of AKT
t_189	_trans_189	[signal_no_death + 1] & [akt + 1] & [AKt_PP - 1];	AKT signaling for survival
t_190	hif_p53_degr	[Ub - 1];	Hif degradation via p53
t_191	_trans_191	[Ub + 1] & [mdm2 + 1] & [HIF_mod - 1] & [mdm2 - 1];	Interaction of Hif with mdm2 for ubiquitination

t_192	_trans_192	[HIF_mod + 1] & [Hif_IN - 1] & [p53 - 1];	Modification of Hif by p53
t_193	p53_degrad	[_place_295 - 1];	P53 degradation by mdm2
t_194	_trans_194	[_place_295 + 1] & [mdm2 + 1] & [mdm2 - 1] & [p53 - 1];	Binding of mdm2 and p53
t_195	_trans_195	[DEATH_signal + 1] & [p300_2 + 1] & [_place_298 - 1];	Signal of p53 for Apoptosis
t_196	_trans_196	[_place_298 + 1] & [p300_2 - 1] & [_place_300 - 1];	Transcription activity of p53
t_197	_trans_197	[_place_300 + 1] & [p53 - 1] & [vhl - 1];	Complex p53 and vhl
t_198	survival	[signal_no_death - 1];	Survival signal by IGF1
t_199	_trans_199	[sp1_solo + 1] & [sp1_P - 1];	Dephosphorylation of Sp1
t_200	Apoptosis	[DEATH_signal - 1];	Apoptosis by p53
t_201	_trans_201	[pkcz2 + 1];	Input for PKCξII
t_202	_trans_202	[vhl + 1];	Input for Vhl
t_203	_trans_203	[hur + 1] & [rec_act + 1] & [vegf_stab - 1] & [_place_19 - 1];	Activation of VEGFR by Vegf stabilized by HuR
t_204	_trans_204	[dna + 1] & [pomc + 1] & [nur77 - 1] & [dna - 1];	Stimulation of production of POMC by Nur77
t_205	_trans_205	[acth + 1] & [pomc - 1];	Transformation of POMC in ACTH
t_206	acth_action	[acth - 1];	Release of ACTH and elimination

t_207	_trans_207	[hur + 1] & [vhl + 1] & [no_stab - 1];	Breaking of complex between HuR and Vhl
t_208	_trans_208	[sp1_solo + 1] & [vhl + 1] & [sp1_vhl - 1];	Breaking of complex between Sp1 and Vhl
t_209	_trans_209	[jade1 + 1];	Input for Jade1
t_210	_trans_210	[jade_instab + 1] & [jade1 - 1];	Instability of Jade1
t_211	jade_degrad	[jade_instab - 1];	Output for Jade1
t_212	_trans_212	[b_catenin + 1];	Input for b-catenin
t_213	_trans_213	[ub_jade - 1];	Degradation of b-catenin
t_214	_trans_214	[proliferation - 1];	Proliferation signal
t_215	_trans_215	[lactic_acid - 1];	Lactic acid elimination
t_216	_trans_216	[pyr + 1];	Input for Pyruvate from other metabolic pathways
t_217	_trans_217	[mmp_inact - 1];	Output for MMP inactive form
t_218	_trans_218	[glicogen - 1];	Glycogen removal
t_219	_trans_219	[tnfa - 1];	Output for TNFα
t_220	_trans_220	[card9 + 1] & [card9_p - 1];	Dephosphorylation of card9
t_221	_trans_221	[ub_moved - 1];	Movement of RPB1 to other areas of nucleus
t_222	_trans_222	[pkcz2 + 1] & [par6 + 1] & [par6_pkcz - 1];	Breaking of complex between par6 and PKCζ
t_223	_trans_223	[tight_j_form - 1];	Tight junction formation
t_224	_trans_224	[H2O - 1];	Output removal

t_225	_trans_225	[p53 + 1];	Input p53
t_226	_trans_226	[rpb7 + 1] & [rpb1 + 1];	Input for RPB1 and RPB7
t_227	_trans_227	[prot_no_ub - 1];	Output for VDU-deubiquitinated proteins
t_228	_trans_228	[prot_ub + 1];	Input for ubiquitinated proteins
t_229	_trans_229	[VDU + 1];	Input for VDU
t_230	_trans_230	[nur77_mRNA + 1] & [vegf2 + 1] & [Hif_IN + 1] & [vhl + 1] & [nur_vhl_hif - 1];	Breaking of complex between Hif, Nur77 and Vhl, stimulation of production of Nur77 and Vegf
t_231	_trans_231	[_place_315 + 1] & [sp1_P + 1] & [sp1_P - 1];	Transcription for IGFR mRNA by sp1
t_232	_trans_232	[_place_315 – 1];	Output for excess mRNA
t_233	_trans_233	[IGF1R -1];	Output for excess IGFR
t_234	_trans_234	[FIH + 1] & [hif + 1] & [O2 + 1] & [Hif_IN - 1] & [FIH - 1] & [O2 – 1];	Inibition HIF by FIH in Normoxia
t_235	_trans_235	[NUR77 + 1];	Input for Nur77
t_236	_trans_236	[_place_318 + 1] & [Cu2 + 1] & [_place_259 - 1] & [Cu2 – 1];	Cu2 interaction with Vhl and PKCζII for degradation.
t_237	_trans_237	[PG2 + 1] & [PGM + 1] & [PG3 - 1] & [PGM – 1];	Modification of 3-Phosphoglycerate to 2-phosphoglycerate

Table S1: **List of model transitions.** *The sequential number, name, biological functions and bibliographic source are listed.*

Place Number	Name	Biological Meaning
p_0	Hif_IN	HIF1α after entering nucleus
p_1	ARNT	Hif1β/ARNT
p_2	HIF_dim	Dimeric form of HIF
p_3	p300	Coactivator p300
p_4	creb	Coactivator CREB
p_5	cjun1	Coactivator cjun
p_6	Creb_300_HIF	HIF-p300-CREB complex
p_7	HIF_act	completely activated form of HIF
p_8	hre	Hypoxia Response Element
p_9	_place_9	DNA transcripted
p_10	_place_10	ET_mRNA
p_11	_place_11	VEGF_mRNA
p_12	_place_12	EPO_mRNA
p_13	_place_13	Lactate DH mRNA
p_14	_place_14	pyrDH Kinase_mRNA
p_15	VEGF	VEGF peptide
p_16	EPO	EPO peptide
p_17	ET1	ET peptide
p_18	pyrDH	PyrDH protein
p_19	_place_19	VEGF receptor
p_20	rec_act	activated VEGF receptor
p_21	_place_21	PLC1 activated
p_22	PI2P	Phosphatydil inositol biphosphate
p_23	DAG1	Diacylglycerol
p_24	_place_24	RAS1activated
p_25	PLC1	Phospholipase C
p_26	ip31	Ip3, 1 to avoid non-unique names
p_27	Ca_chan1	calcium Channel
p_28	ca1	Calcium Ion
p_29	cam_k_kk1	Calmodulin Kinase system
p_30	PKC1	Protein Kinase C
p_31	_place_31	PKC1 activated
p_32	_place_32	VEGF signal pathway 3 effects
p_33	_place_33	VEGF signal pathway 3 effects
p_34	PI3K	Phosphoinositide 3 kinase
p_35	PI3k_act	Phosphoinositide 3 kinase-active state
p_36	PIP3	Phosphoinositol 3 phosphate
p_37	PDK	Phosphoinositide dependent kinase
p_38	PKB	Protein Kinasi B also known as Akt
p_39	PKB_PP	PKB phosphorilated
p_40	pdk_PP	PDK phosphorilated
p_41	BAD	B Apoptosis Domain
p_42	BAP_DIP	complex BAD and DIP
p_43	_place_43	DIP after action

Place Number	Name	Biological Meaning
p_44	inhib	BAD inhibitor protein
p_45	bad_inact	BAD inactive state
p_46	DIP	DIP - PBAP Chromatin Remodeling protein
p_47	_place_47	Vegf signal pathway 2 effects
p_48	RAS1	RAS GTPase
p_49	_place_49	Mapkkk inactive
p_50	mapkkk	mapkkk active
p_51	_place_51	mapkk inactive
p_52	mapkk	mapkk active
p_53	_place_53	mapk inactive
p_54	mapk	active mapk
p_55	gdp1	Guanosine Diphosphate
p_56	gtp1	Guanosine Triphosphate
p_57	ATP	Adenosine Triphosphate
p_58	ADP	Adenosine Diphosphate
p_59	_place_59	Vegf signal pathway 1 effects
p_60	pkb_act	Protein Kinase B – active state
p_61	_place_61	Inactive EPO receptor
p_62	EPOR	active EPO Receptor
p_63	_place_63	inactive PLC
p_64	PLC_ACT	Phospholipase C – active state
p_65	stat5	transcriptor factor
p_66	jak2	transcriptor factor
p_67	jak	transcriptor factor
p_68	STATx2	dimer form of stat5
p_69	shc2	Src homology containing protein
p_70	_place_70	active shc
p_71	SOS12	Son of Sevenless protein
p_72	GRB2	Growth factor receptor-bound protein 2
p_73	complex1	Shc-sos-grb complex
p_74	GTP2	Guanosine Triphosphate
p_75	RAS3	RAS GTP-ase
p_76	_place_76	inactive RAF2, serine/threonine kinase
p_77	RAF2	RAF2 serine/threonine kinase
p_78	_place_78	inactive mek1
p_79	MEk1	Mitogen-activated protein kinase kinase
p_80	_place_80	inactive ERK1
p_81	ERK1	Extracellular signal-regulated kinases
p_82	_place_82	inactive RAS3
p_83	_place_83	GDP - Guanosine diphosphate
p_84	_place_84	ERK in nucleus
p_85	ELK1	ETS domain-containing protein
p_86	_place_86	ELK1 inactive
p_87	_place_87	DNA
p_88	_place_88	CFOS mRNA
p_89	_place_89	EPO effects1
p_90	CFOS	Transcription factor
p_91	cjun	Transcription factor

Place Number	Name	Biological Meaning
p_92	_place_92	Cfos, cjun complex
p_93	_place_93	inactive cjun
p_94	_place_94	EPO effects2
p_95	_place_95	cjun activating protein
p_96	ETRa	Endothelin Receptor a
p_97	_place_97	ETRa active
p_98	_place_98	ETRa2 active
p_99	ETRb	Endothelin Receptor b
p_100	_place_100	Protein G - active
p_101	_place_101	shc active state
p_102	PG	Protein G
p_103	shc1	Src homology containing protein
p_104	PLC2	Phospholipase C
p_105	_place_105	PLC2 active
p_106	pip	Phosphatinositol phosphate
p_107	DAG	diacylglycerol
p_108	IP3	Inositol triphosphate
p_109	PKC2	Protein kinase C
p_110	_place_110	PKC2 active state
p_111	_place_111	Ras2 active state
p_112	RAs2	Ras
p_113	raf2	RAF2 serine/threonine kinase
p_114	_place_114	raf2 active state
p_115	mapkk2	Mitosis activated protein kinase kinase
p_116	_place_116	Mapkk2 active state
p_117	mapk2	Mitosis activated protein kinase
p_118	_place_118	mapk2 active state
p_119	ca	Calcium ion
p_120	cam_k_kk2	Calcium-calmodulin-kinase complex
p_121	camk_pp	Calmodulin kinase complex phosphorilated state
p_122	CREB_P	CREB phosphorilated
p_123	_place_123	CREB
p_124	PI3k	Phosphoinositide 3 kinase
p_125	csrc1	SRC
p_126	sos1	Son of sevenless
p_127	grb1	GRB
p_128	complex	Sos-shc-grb complex
p_129	glu1	glucose
p_130	_place_130	glucose inside cells
p_131	G6P	Glucose 6 Phosphate
p_132	F6P	Fructose 6 phosphate
p_133	F16BP	Fructose 1,6- diphosphate
p_134	DHAP	duhydroxyacetone phosphate
p_135	GADP	D-glyceraldehyde 3 phosphate
p_136	exokinase	Hexokinase enzime
p_137	Isomerase	Glucose-6-phosphate isomerase
p_138	PFK	Phopshofructokinase
p_139	ALDO	Aldolase

Place Number	Name	Biological Meaning
p_140	TPI	triosephosphato isomerase
p_141	GAPDH	glyceraldehyde phosphate dehydrogenase
p_142	P	Pyrophosphate
p_143	NAD	
p_144	NADH	
p_145	BPG	Biphosphoglycerate
p_146	PG3	3phosphoglycerate
p_147	ENO	
p_148	pep	Phophoenolpyruvate
p_149	Pyr_kin	Pyruvate kinase
p_150	pyr	Pyruvate
p_151	lactic_acid	
p_152	_place_152	Pyruvate carboxylase
p_153	LDH	Lactate dehydrogenase
p_154	COA	Coenzyme A
p_155	AChCOA	Acetyl coenzyme A
p_156	_place_156	active pyrdh
p_157	_place_157	inactive pyr dh
p_158	oxa	Oxalacetate
p_159	cit_synt	citrate s
p_160	Cyt	Cytrate
p_161	cisAco	Cis-aconitate
p_162	aconitase	
p_163	isocyt	Isocytrate
p_164	oxasu	oxalosuccinate
p_165	isocyt_DH	isocitrate dehydrogenase
p_166	a_cheto	A-ketoglutarate
p_167	acheto_DH	A-ketoglutarate dehydrogenase
p_168	SuCoA	Succinil-coA
p_169	SuCoA_synth	Succinil-coA synthetase
p_170	su	Succinate
p_171	su_DH	Succinate Dehydrogenase
p_172	fum	Fumarate
p_173	mal	Malate
p_174	fumarase	
p_175	glut2	GLUT2
p_176	_place_176	glucose inside pancreas cells
p_177	insulin	
p_178	glicogen	
p_179	H2O	water
p_180	_place_180	phosphatase inactivating GS
p_181	GS_inact	inactive Glycogen synthetase
p_182	GS_act	active Glycogen synthetase
p_183	PKB_AKT	Protein Kinase B aka AKT
p_184	PKB_akt	inactive PKB
p_185	gsk3b_inact	Glycogen Synthetase Kinase 3b inactive
p_186	gsk3b_act	active glycogen synthetase kinase 3b

Place Number	Name	Biological Meaning
p_187	_place_187	pancreatic insulin synthesis complex
p_188	_place_188	active insulin synthesis system
p_189	glu2	Glucose inside muscles and fat cells
p_190	hif_ppp	Phosphorylated HIF
p_191	ub_ase	Ubiquitinating system
p_192	hif_ppp_ub	Phosphorylated and ubiquitinated hif
p_193	vhl	Von Hippel Lindau
p_194	vhl_pp	Phosphorylated VHL
p_195	_place_195	PHD2 inactivated by Krebs cycle analogues
p_196	ARD	Aryl Ribonuclease domain
p_197	_place_197	active ARD
p_198	PHD2_act	active PHD2
p_199	O2	oxygen
p_200	Hy_Ac_HIF1a	Modified HIF
p_201	eloc	Elongin C
p_202	elob	Elongin B
p_203	vcd	complex of VHL elongins and Cu2-ring box
p_204	Hif_ub	HIF after ubiquitination
p_205	_place_205	inhibiting analogues of oxoglutarate for PHD2
p_206	p2	phosphate
p_207	GDP2	Guanosine diphosphate
p_208	GTP3	Guanosine triphosphate
p_209	P_ase	Phosphatase
p_210	mt	Microtubules
p_211	tubulin	Tubulin
p_212	nedd8	Nedd8
p_213	nedd_vhl	Neddylated VHL
p_214	FN	Fibronectine
p_215	VHL_FN	complex of Vhl with Fibronectine
p_216	stable_matrix	crossed linked matrix
p_217	MT1MMP	Metallo proteinase 1
p_218	_place_218	metallo proteinase mRNA
p_219	mmp_inact	inactivated MMP
p_220	timp1	Tissue inhibiting Metallo proteinase1
p_221	_place_221	inactive Tissue inhibiting metallo proteinase1
p_222	prot_no_ub	deubiquitinated generic protein
p_223	prot_ub	ubiquitinated protein
p_224	Cu2	cullin2
p_225	_place_225	complex VDU and vcb
p_226	VDU	VHL interactiong deubiquitinase
p_227	_place_227	Ubiquitinated VDU
p_228	_place_228	Vhl moved out of nucleus
p_229	eEF1a	Endonuclear export farctor1a
p_230	_place_230	Functional Rna polimerase
p_231	_place_231	rna polimarase other subunits
p_232	rpb1	Rna polimerasi binding protein *subunit1
p_233	card9	Caspase associated recruiting domain

Place Number	Name	Biological Meaning
p_234	nfkb	NF-kB
p_235	_place_235	activated nfkb+card9
p_236	_place_236	transcription activity
p_237	tnfa	tumor necrosis factor a
p_238	_place_238	complex VCB card9
p_239	ck2	Kinase activity on Card9
p_240	_place_240	ck2 inactive
p_241	card9_p	card9 phosphorylated
p_242	_place_242	Rbp-vhl
p_243	phy_rpb1	Prolin-hydroxylated RPB1
p_244	rpb_kin	Rpb kinase
p_245	Rpb_hy_p	Rpb- hydroxilated and phosphorilated form
p_246	ub	ubiquitin
p_247	ub_moved	rpb1-hydroxphosphorilated-ubiquitinated and moved out of nucleus
p_248	Rpb7_vhl	Rpb subunit 7 in complex with vhl
p_249	GSK3B_complex	Glicogen sintase kinase 3 b complex with
p_250	stab_complex	Stable jade complex
p_251	jade1	jade
p_252	proliferation	effect
p_253	b_catenin	Beta catenin
p_254	b_cat_P	Beta catenin phosphorilated form
p_255	ub_jade	Jade ubiquitinated
p_256	wnt	Signaling Pathways: Wnt / β-Catenin Signaling
p_257	_place_257	APC
p_258	_place_258	Axin
p_259	_place_259	VHL in complex with PKCzII
p_260	pkcz2	Atipical Protein kinase C zeta II
p_261	par6	Subunit of tight junction par 6
p_262	aPKC	Atipical protein kinase C unknown type
p_263	_place_263	Par6 in complex with aPKC
p_264	cdc42	CDC42 is a protein involved in regulation of the cell cycle
p_265	aPKC_act	Atypical PKC unknown type- active state
p_266	par3	Subunit of tight junction par 3
p_267	par3_P	Par 3 phosphorilated form
p_268	tight_j_form	Formation of tight junction
p_269	par6_pkcz	Complex of pkcz with par 6
p_270	nur77_mRNA	MRNA of protein nur77 nuclear receptor
p_271	sp1_vhl	Sp1 transcriptor factor in complex with vhl
p_272	sp1_P	Sp1 phosphorilated form
p_273	vegf_stab	Stable VEGF
p_274	vegf2	Additional vegf to increase action of HIF-induced
p_275	nur_vhl_hif	Complex nur 77 with HIF and VHL
p_276	no_stab	instability
p_277	nur_vhl	Complex of vhl and nur77
p_278	nur77	Nuclear receptor family 4 type 1a
p_279	hur	hur RNA-binding protein
p_280	stabilization	Stabilization of IGFR mRNA

Place Number	Name	Biological Meaning
p_281	IGF1R	Insulin like growth factor receptor
p_282	IGF1	Insulin like growth factor 1
p_283	igfr_act	Insulin like growth factor receptor – active state
p_284	pi3k	Phosphoinositide 3 kinase
p_285	pi3k_act	Phosphoinositide 3 kinase – active state
p_286	pip3	Phosphatinositol phosphate
p_287	akt	PKB - Akt, also known as Protein Kinase B (PKB)
p_288	AKT_mem	PKB activated with PIP3
p_289	mtorc2_1	Mtorc protein
p_290	PDPK1_2	Prolin Directed protein Kinase
p_291	signal_no_death	Signal for survival
p_292	AKt_PP	PKB phosphorylated
p_293	Ub	ubiquitin
p_294	HIF_mod	Modified Hif by p53
p_295	_place_295	P53 and Mdm2, towards degradation
p_296	mdm2	MDM2binding domain
p_297	DEATH_signal	Signal for apoptosis
p_298	_place_298	p300 and p53 complex
p_299	p300_2	Coactivator of transcription p300
p_300	_place_300	p53 and vhl complex
p_301	p53	Tumorsuppressor p53
p_302	rpb7	Rna polimerase subunit 7
p_303	sp1_solo	Sp1 alone
p_304	hif	Hipoxia inducible factor 1a
p_305	dna	Desoxy ribonucleic acid
p_306	pomc	Proopiomelanocortin
p_307	acth	Adenocorticotropic Hormone
p_308	jade_instab	Jade – unstable form
p_309	_place_309	bcatenin activated by wnt
p_310	PHD2	Prolin Hydroxilating Domain containing Protein 2
p_311	Et_eff1	Endotelin effects 1 vasocostriction
p_312	Et_eff2	Endotelin effects 2 vasocostriction high blood pressure
p_313	_place_313	DNA in et1
p_314	_place_314	ca channel in et
p_315	_place_315	IGFR mRNA
p_316	mal_DH	Malate dehydrogenase
p_317	FIH	Factor inhibiting hif
p_318	_place_318	ubiquitinated pkcz2
p_319	PG2	2-phosphoglycerate
p_320	PGM	Phosphoglycerate mutase
p_321	PGK	Phophoglycerate kinase
p_322	vcb	VHL in complex with elongins B and C

Table S2: List of all model places. The progressive ID number, name and biological meaning are shown.

# Inv	Transitions
Inv_60	1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 15, 16, 17, 18, 19, 20, 22, 28, 29, 30, 32, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68.absorbed, 69.eating, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 82, 83, 97, 98, 113, 114, 115, 116, 117, 118, 119.degradation1, 123, 124, 125, 127, 128, 129, 134, 135, 136.matrix, 137, 138, 142, 215, 224, 237
Inv_87	1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 15, 22, 28, 29, 30, 32, 33, 34, 37, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68.absorbed, 69.eating, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 82, 83, 97, 98, 113, 114, 115, 116, 117, 118, 119.degradation1, 123, 125, 129, 130, 134, 135, 136.matrix, 137, 138, 142, 175, 176, 178, 180, 203, 215, 224, 230, 237,
Inv_88	1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 15, 22, 28, 29, 30, 32, 35, 36, 37, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68.absorbed, 69.eating, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 82, 83, 97, 98, 113, 114, , 115, 116, 117, 118, 119.degradation1, 123, 125, 129, 130, 134, 135, 136.matrix, 137, 138, 142, 175, 176, 178, 180, 203, 215, 224, 230, 237
Inv_89	1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 15, 22, 28, 29, 30, 32, 33, 34, 37, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68.absorbed, 69.eating, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 82, 83, 97, 98, 113, 114, 115, 116, 117, 118, 119.degradation1, 123, 125, 129, 130, 138, 139, 140, 142, 175, 176, 178, 180, 202, 203, 215, 217, 224, 230, 237
Inv_90	1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 15, 22, 28, 29, 30, 32, 35, 36, 37, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68.absorbed, 69.eating, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 82, 83, 97, 98, 113, 114, 115, 116, 117, 118, 119.degradation1, 123, 125, 129, 130, 138, 139, 140, 142, 175, 176, 178, 180, 202, 203, 215, 217, 224, 230, 237
Inv_91	1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 15, 22, 28, 29, 30, 32, 33, 34, 37, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68.absorbed, 69.eating, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 82, 83, 97, 98, 106, 107, 108.degradation2, 113, 114, 115, 116, 117, 118, 119.degradation1, 123, 125,

# Inv	Transitions
	129, 130, 138, 139, 140, 142, 175, 176, 178, 180, 203, 215, 217, 224, 230, 237
Inv_92	1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 15, 22, 28, 29, 30, 32, 35, 36, 37, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68.absorbed, 69.eating, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 82, 83, 97, 98, 106, 107, 108.degradation2, 113, 114, 115, 116, 117, 118, 119.degradation1, 123, 125, 129, 130, 138, 139, 140, 142, 175, 176, 178, 180, 203, 215, 217, 224, 230, 237
Inv_93	0, 1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 15, 22, 28, 29, 30, 32, 33, 34, 37, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68.absorbed, 69.eating, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 82, 83, 97, 98, 113, 114, 115, 116, 117, 118, 119.degradation1, 123, 125, 129, 130, 138, 139, 140, 142, 175, 176, 178, 180, 190.hif_p53_degr, 191, 192, 203, 215, 217, 224, 225, 230, 237,
Inv_94	0, 1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 15, 22, 28, 29, 30, 32, 35, 36, 37, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68.absorbed, 69.eating, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 82, 83, 97, 98, 113, 114, 115, 116, 117, 118, 119.degradation1, 123, 125, 129, 130, 138, 139, 140, 142, 175, 176, 178, 180, 190.hif_p53_degr, 191, 192, 203, 215, 217, 224, 225, 230, 237
Inv_101	1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 15, 22, 28, 29, 30, 32, 33, 34, 37, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 79, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 97, 98, 113, 114, 115, 116, 117, 118, 119.degradation1, 123, 125, 129, 130, 131, 134, 135, 136.matrix, 137, 138, 142, 215, 216, 224
Inv_105	1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 15, 22, 28, 29, 30, 32, 35, 36, 37, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 79, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 97, 98, 113, 114, 115, 116, 117, 118, 119.degradation1, 123, 125, 129, 130, 131, 134, 135, 136.matrix, 137, 138, 142, 215, 216, 224
Inv_125	1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 16, 17, 18, 19, 20, 32, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68.absorbed, 69.eating, , 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 82, 83, 97, 98, 113, 114, 115, 116, 117, 118, 119.degradation1, 123, 124, 127, 128, 129, 138, 139, 140, 202, 215, 216, 217, 224, 237
Inv_129	1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 15, 16, 17, 18, 19, 20, 22, 28, 29, 30, 32, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68.absorbed, 69.eating, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 82, 83, 97, 98, 113, 114, 115, 116, 117, 118, 119.degradation1, 123, 124, 125, 127, 128, 129, 138, 139, 140, 142, 202, 215, 217, 224, 237
Inv_142	1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 14, 15, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68.absorbed, 69.eating, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 82, 83, 97, 98, 113, 114, 115, 116, 117, 118, 119.degradation1, 123, 125, 126, 127, 129, 138, 139, 140, 142, 202, 215, 217, 224, 237
Inv_144	1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 15, 22, 28, 29, 30, 32, 33, 34, 37, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 79, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 97, 98, 113, 114, 115, 116, 117, 118, 119.degradation1, 123, 125, 129, 130, 131, 138, 139, 140, 142, 202, 215, 216, 217, 224
Inv_148	1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 15, 22, 28, 29, 30, 32, 35, 36, 37, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 79, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 97, 98, 113, 114, 115, 116, 117, 118, 119.degradation1, 123, 125, 129, 130, 131, 138, 139, 140, 142, 202, 215, 216, 217, 224
Inv_172	1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 15, 16, 17, 18, 19, 20, 22, 28, 29, 30, 32, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68.absorbed, 69.eating, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 82, 83, 97, 98, 106, 107, 108.degradation2, 113, 114, 115, 116, 117, 118, 119.degradation1, 123, 124, 125, 127, 128, 129, 138, 139, 140, 142, 215, 217, 224, 237
Inv_185	1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 14, 15, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68.absorbed, 69.eating, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 82, 83, 97, 98, 106, 107, 108.degradation2, 113, 114, 115, 116, 117, 118, 119.degradation1, 123, 125, 126, 127, 129, 138,

# Inv	Transitions
	139, 140, 142, 215, 217, 224, 237
Inv_215	0, 1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 15, 16, 17, 18, 19, 20, 22, 28, 29, 30, 32, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68.absorbed, 69.eating, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 82, 83, 97, 98, 113, 114, 115, 116, 117, 118, 119.degradation1, 123, 124, 125, 127, 128, 129, 138, 139, 140, 142, 190.hif_p53_degr, 191, 192, 215, 217, 224, 225, 237
Inv_227	0, 1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 14, 21, 23, 24, 25, 26, 27, 31, 32, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 97, 98, 111, 112, 113, 114, 115, 116, 117, 118, 119.degradation1, 121, 123, 126, 127, 129, 131, 138, 139, 140, 190.hif_p53_degr, 191, 192, 215, 216, 217, 224, 225
Inv_228	0, 1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 14, 15, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68.absorbed, 69.eating, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 82, 83, 97, 98, 113, 114, 115, 116, 117, 118, 119.degradation1, 123, 125, 126, 127, 129, 138, 139, 140, 142, 190.hif_p53_degr, 191, 192, 215, 217, 224, 225, 237
Inv_245	0, 1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 15, 22, 28, 29, 30, 32, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68.absorbed, 69.eating, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 97, 98, 111, 112, 113, 114, 115, 116, 117, 118, 119.degradation1, 121, 123, 125, 127, 129, 131, 138, 139, 140, 142, 190.hif_p53_degr, 191, 192, 215, 217, 224, 225, 237
Inv_278	0, 1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 14, 21, 23, 24, 25, 26, 27, 31, 32, 33, 34, 37, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 97, 98, 111, 112, 113, 115, 116, 117, 118, 119.degradation1, 120, 123, 126, 129, 130, 131, 138, 139, 140, 190.hif_p53_degr, 191, 192, 215, 216, 217, 224, 225
Inv_279	0, 1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 14, 21, 23, 24, 25, 26, 27, 31, 32, 35, 36, 37, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 97, 98, 111, 112, 113, 115, 116, 117, 118, 119.degradation1, 120, 123, 126, 129, 130, 131, 138, 139, 140, 190.hif_p53_degr, 191, 192, 215, 216, 217, 224, 225
Inv_280	0, 1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 14, 21, 23, 24, 25, 26, 27, 31, 32, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 97, 98, 111, 112, 113, 115, 116, 117, 118, 119.degradation1, 120, 123, 126, 127, 129, 131, 138, 139, 140, 190.hif_p53_degr, 191, 192, 215, 216, 217, 224, 225
Inv_377	0, 1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 14, 21, 23, 24, 25, 26, 27, 31, 32, 33, 34, 37, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 97, 98, 111, 112, 113, 115, 116, 117, 118, 119.degradation1, 120, 122, 123, 126, 129, 130, 131, 138, 139, 140, 190.hif_p53_degr, 191, 192, 215, 216, 217, 224, 225

Table S3: List of T-invariants named in the text and their composition.

C8 (72)	Inv_280, Inv_384, Inv_227, Inv_383, Inv_226, Inv_225, Inv_271, Inv_357, Inv_184, Inv_356, Inv_183, Inv_262, Inv_330, Inv_141, Inv_329, Inv_140, Inv_182, Inv_139, Inv_251, Inv_301, Inv_72, Inv_300, Inv_71, Inv_70, Inv_283, Inv_392, Inv_239, Inv_391, Inv_238, Inv_237, Inv_274, Inv_365, Inv_196, Inv_364, Inv_195, Inv_265, Inv_338, Inv_153, Inv_337, Inv_152, Inv_194, Inv_151, Inv_256, Inv_311, Inv_110, Inv_310, Inv_109, Inv_108, Inv_277, Inv_375, Inv_214, Inv_374, Inv_213, Inv_212, Inv_268, Inv_348, Inv_171, Inv_347, Inv_170, Inv_259, Inv_321, Inv_128, Inv_320, Inv_127, Inv_169, Inv_126, Inv_248, Inv_292, Inv_59, Inv_291, Inv_58, Inv_57
C9 (144)	Inv_381, Inv_223, Inv_380, Inv_222, Inv_279, Inv_278, Inv_378, Inv_219, Inv_377, Inv_218, Inv_221, Inv_217, Inv_270, Inv_269, Inv_261, Inv_260, Inv_354, Inv_180, Inv_353, Inv_179, Inv_351, Inv_176, Inv_350, Inv_175, Inv_327, Inv_137, Inv_326, Inv_136, Inv_324, Inv_133, Inv_323, Inv_132, Inv_178, Inv_174, Inv_135, Inv_131, Inv_298, Inv_68, Inv_297, Inv_67, Inv_250, Inv_249, Inv_295, Inv_64, Inv_294, Inv_63, Inv_66, Inv_62, Inv_390, Inv_236, Inv_389, Inv_235, Inv_282, Inv_281, Inv_387, Inv_232, Inv_386, Inv_231, Inv_234, Inv_230, Inv_273, Inv_272, Inv_264, Inv_263, Inv_363, Inv_193, Inv_362, Inv_192, Inv_360, Inv_189, Inv_359, Inv_188, Inv_336, Inv_150, Inv_335, Inv_149, Inv_333, Inv_146, Inv_332, Inv_145, Inv_191, Inv_187, Inv_148, Inv_144, Inv_309, Inv_107, Inv_308, Inv_106, Inv_255, Inv_254, Inv_306, Inv_103, Inv_305, Inv_102, Inv_105, Inv_101, Inv_372, Inv_210, Inv_371, Inv_209, Inv_276, Inv_275, Inv_369, Inv_206, Inv_368, Inv_205, Inv_208, Inv_204, Inv_267, Inv_266, Inv_258, Inv_257, Inv_345, Inv_167, Inv_344, Inv_166, Inv_342, Inv_163, Inv_341, Inv_162, Inv_318, Inv_124, Inv_317, Inv_123, Inv_315, Inv_120, Inv_314, Inv_119, Inv_165, Inv_161, Inv_122, Inv_118, Inv_289, Inv_55, Inv_288, Inv_54, Inv_247, Inv_246, Inv_286, Inv_51, Inv_285, Inv_50, Inv_53, Inv_49
C10 (52)	Inv_228, Inv_185, Inv_142, Inv_73, Inv_382, Inv_224, Inv_355, Inv_328, Inv_181, Inv_138, Inv_299, Inv_69, Inv_245, Inv_244, Inv_242, Inv_202, Inv_201, Inv_159, Inv_158, Inv_199, Inv_156, Inv_116, Inv_115, Inv_113, Inv_393, Inv_243, Inv_366, Inv_200, Inv_339, Inv_157, Inv_241, Inv_240, Inv_198, Inv_197, Inv_155, Inv_154, Inv_312, Inv_114, Inv_112, Inv_111, Inv_215, Inv_172, Inv_129, Inv_60, Inv_373, Inv_211, Inv_346, Inv_319, Inv_168, Inv_125, Inv_290, Inv_56
C11 (64)	Inv_388, Inv_385, Inv_361, Inv_358, Inv_334, Inv_331, Inv_94, Inv_93, Inv_233, Inv_229, Inv_81, Inv_80, Inv_92, Inv_91, Inv_90, Inv_89, Inv_190, Inv_186, Inv_147, Inv_143, Inv_79, Inv_78, Inv_77, Inv_76, Inv_307, Inv_304, Inv_88, Inv_87, Inv_104, Inv_100, Inv_75, Inv_74, Inv_379, Inv_376, Inv_220, Inv_216, Inv_352, Inv_349, Inv_325, Inv_322, Inv_177, Inv_173, Inv_134, Inv_130, Inv_296, Inv_293, Inv_65, Inv_61, Inv_370, Inv_367, Inv_207, Inv_203, Inv_343, Inv_340, Inv_316, Inv_313, Inv_164, Inv_160, Inv_121, Inv_117, Inv_287, Inv_284, Inv_52, Inv_48

Table S4: Composition in terms of invariants of the clusters C8, C9, C10, C11.

8.3 Evaluation of the steric impact of flavin adenine dinucleotide in Drosophila melanogaster cryptochrome function. Supplementary Information.

*Supplementary Material 1: RING network of the most conserved residues in cryptochromes.*

Supplementary Material 2: Root mean square deviation (RMSD) plots of molecular dynamics simulations: A – holoprotein, B – Cryptochrome without FAD, C – Cryptochrome withtout both FAD and Mg^{2+}, D – Cryptochrome with phosphorylated S526.

Supplementary Material 3: Root mean square fluctuation (RMSF) plots of molecular dynamics simulations: A – holoprotein, B – Cryptochrome without FAD, C – Cryptochrome withtout both FAD and Mg^{2+}, D – Cryptochrome with phosphorylated S526.

Supplementary Material 4: BLUUES surface of A – PL and B – dCRY

Supplementary Material 3
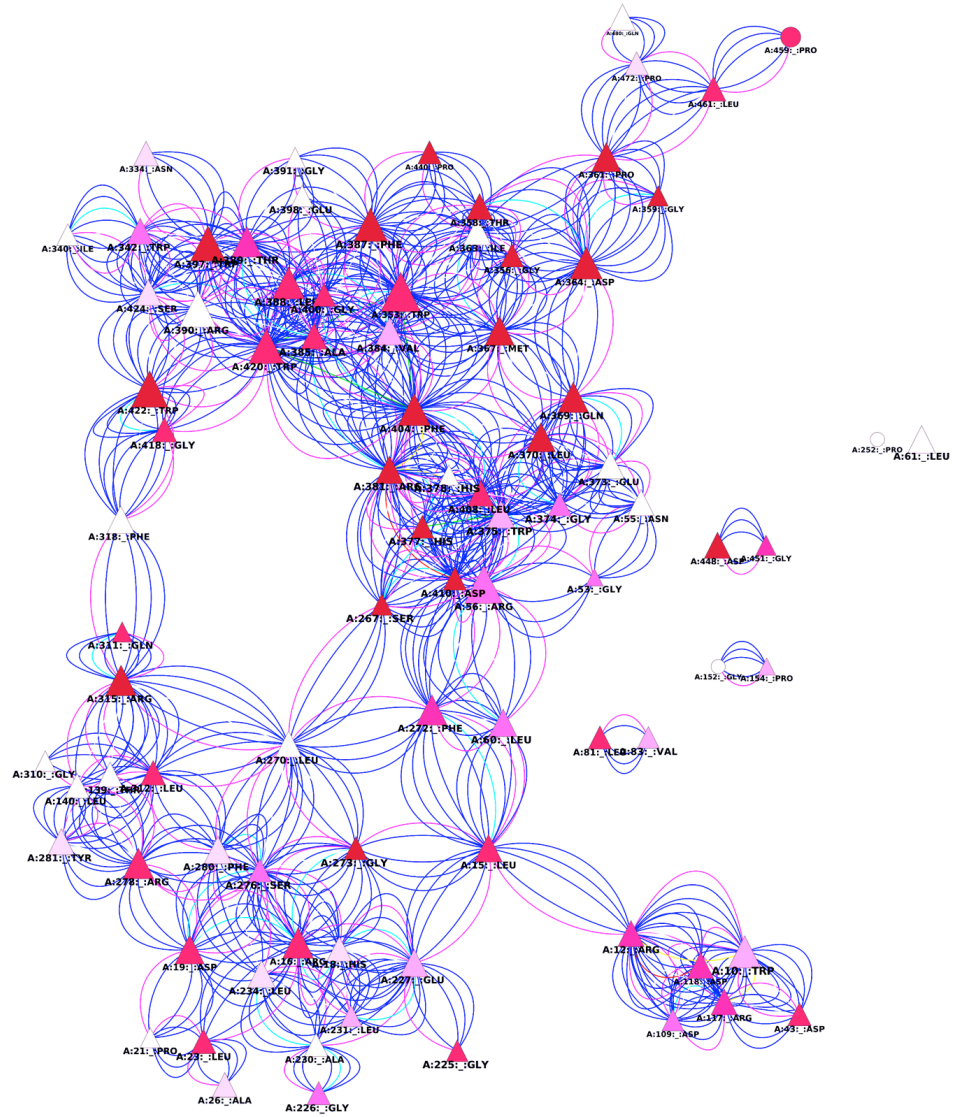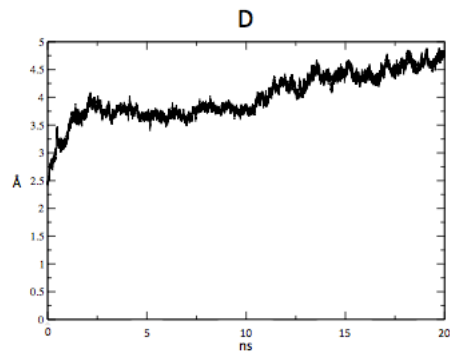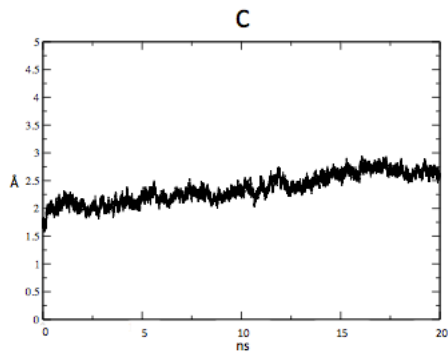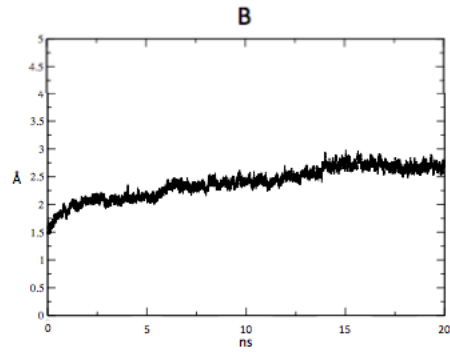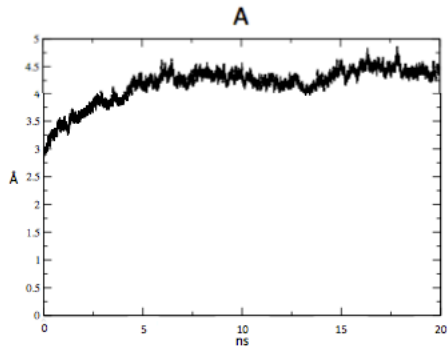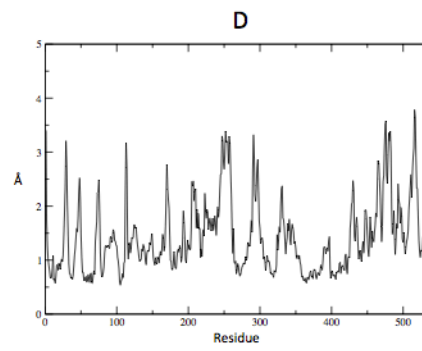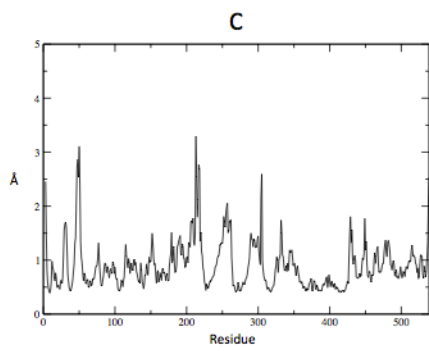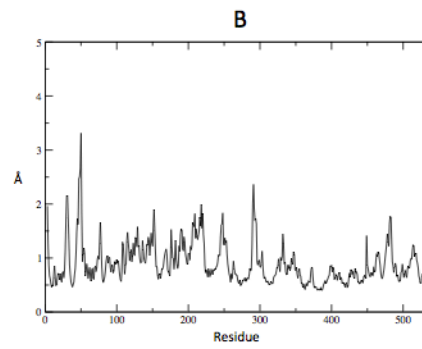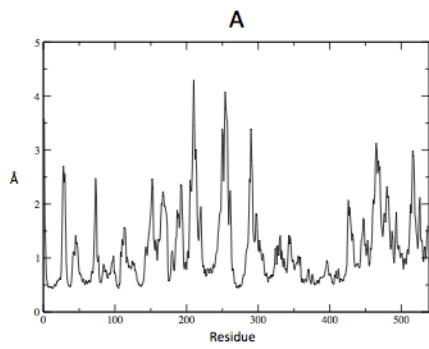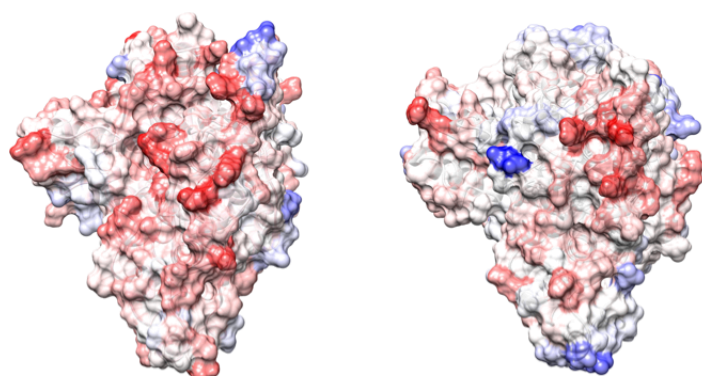
8.4 Structural protein reorganization and fold emergence investigated through amino acid sequence permutations. Supplementary Information.
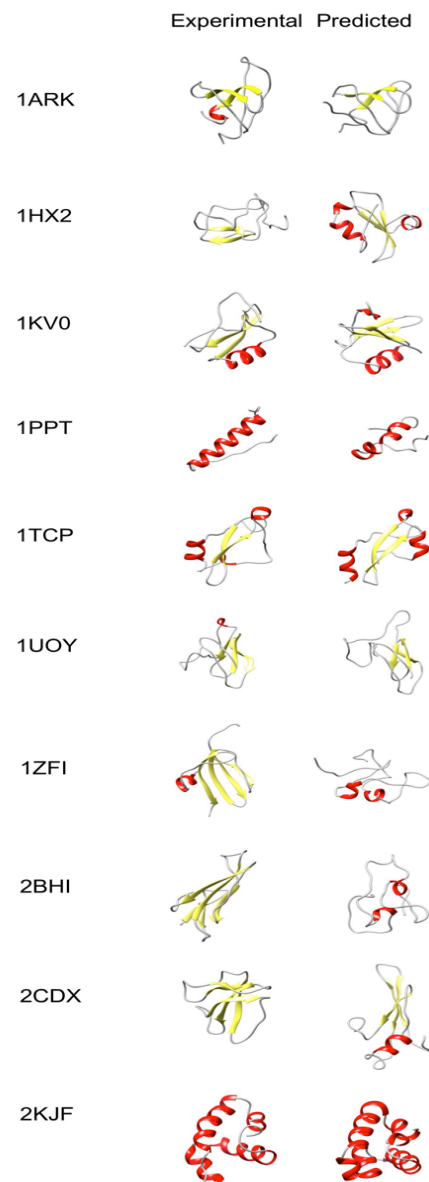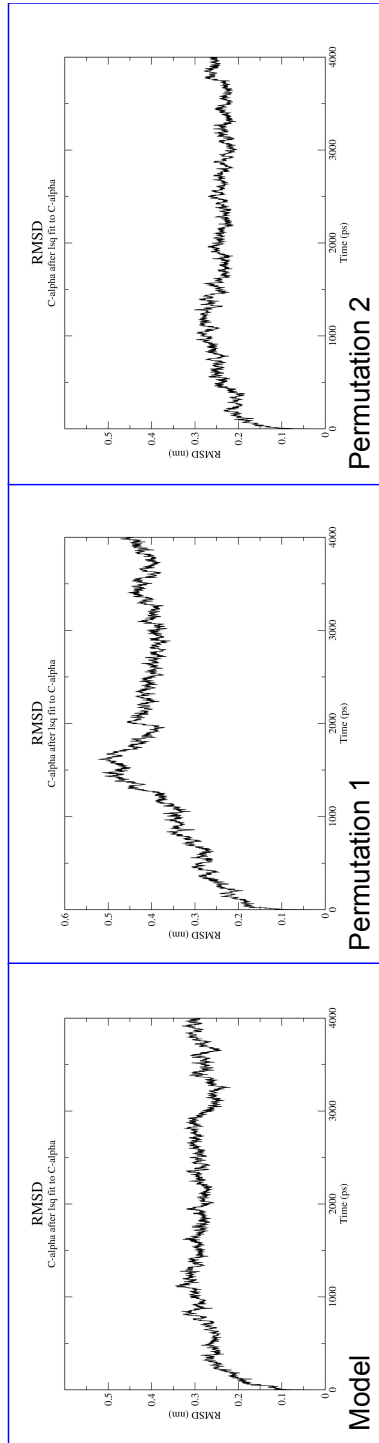

Supplementary Figure 1: Validation of ab initio protocol. Graphical representation of structures obtained with the ab initio method and the corresponding crystal structure
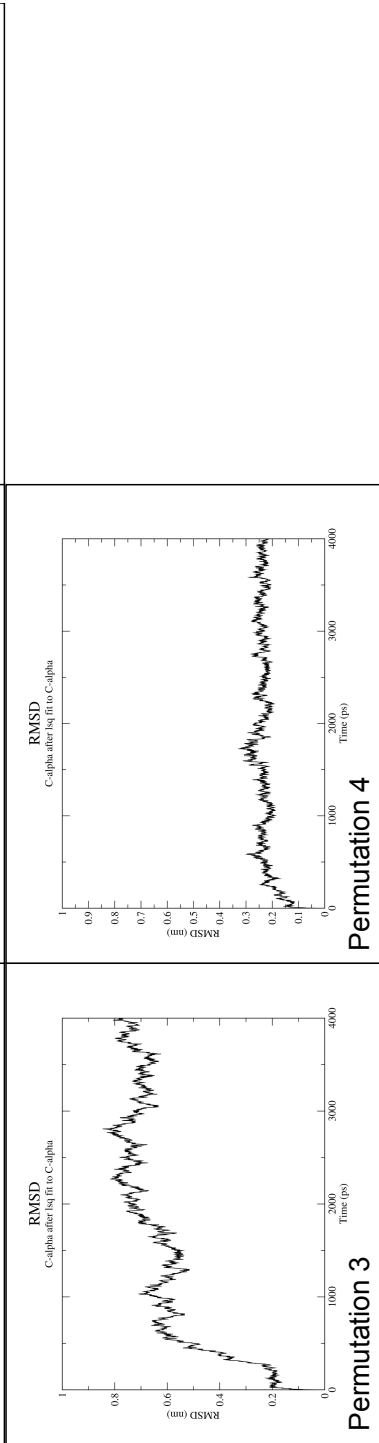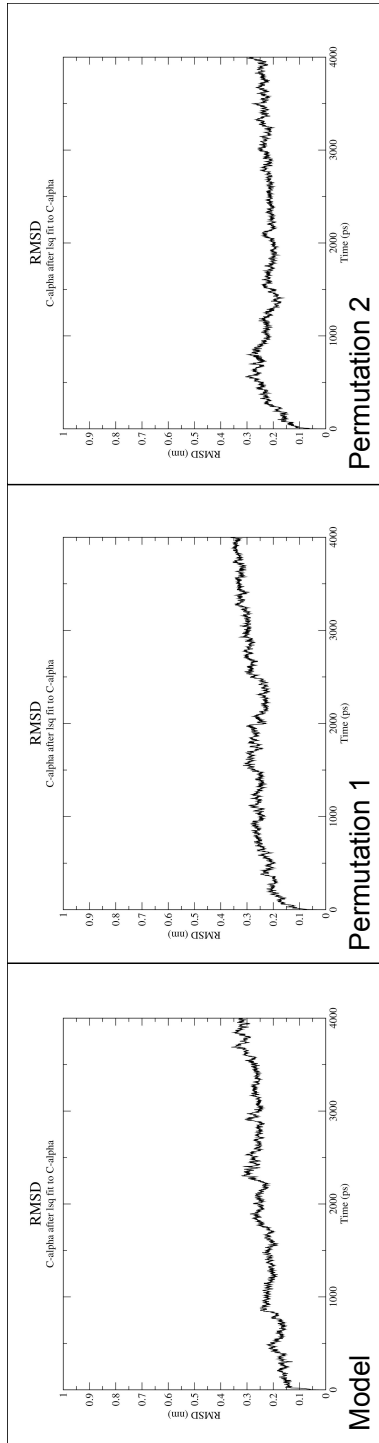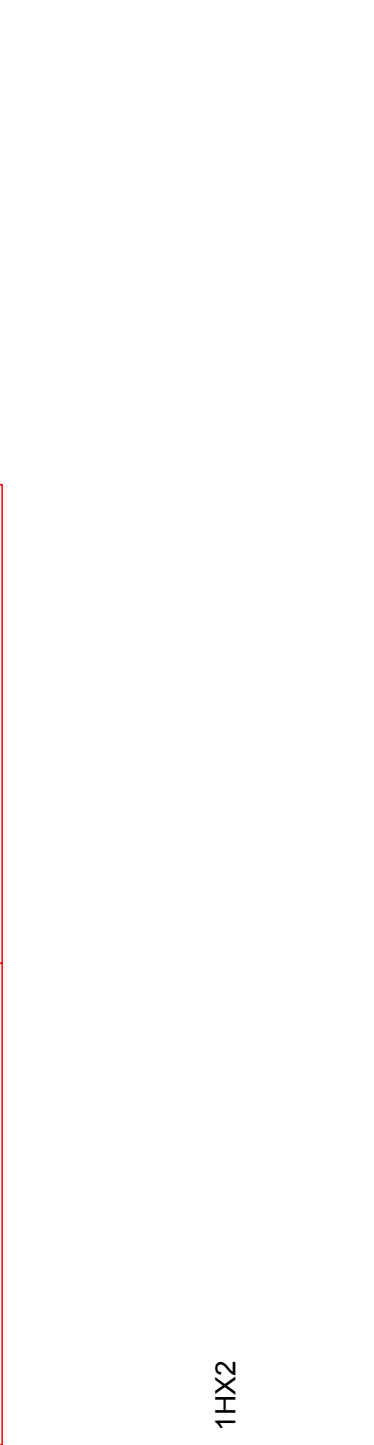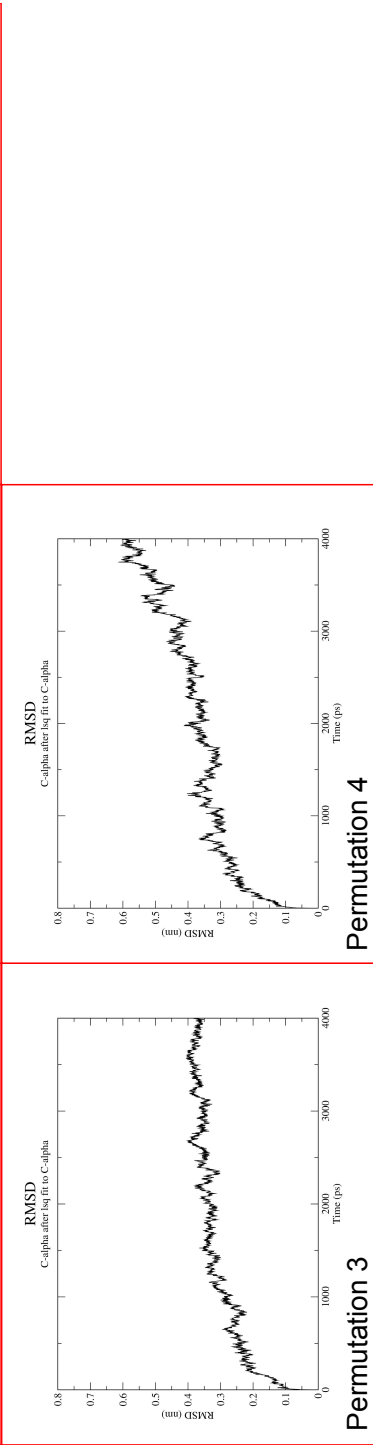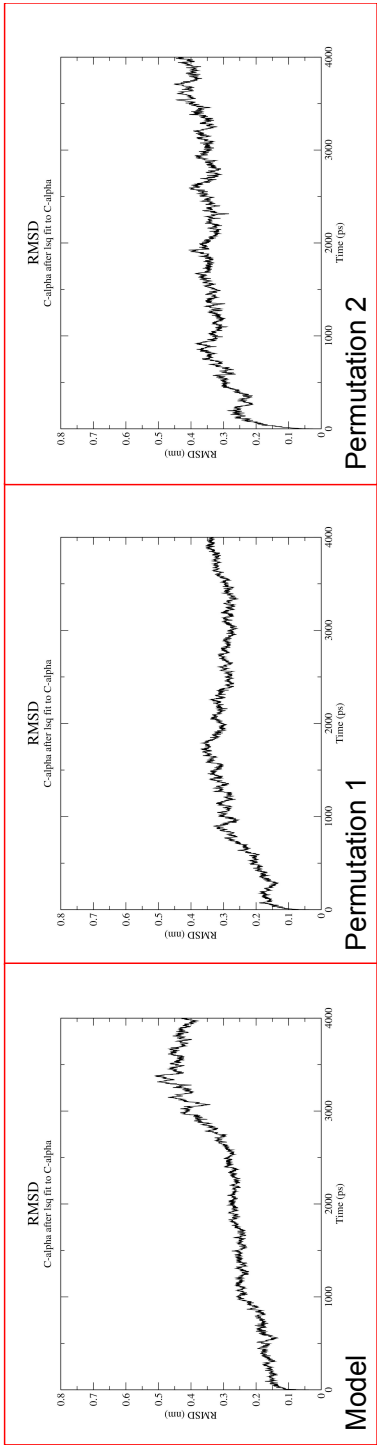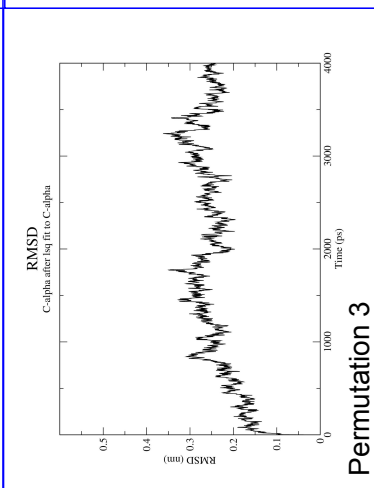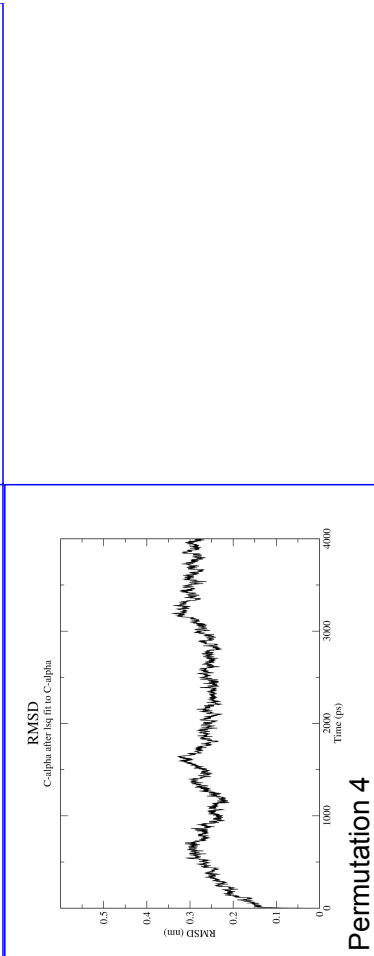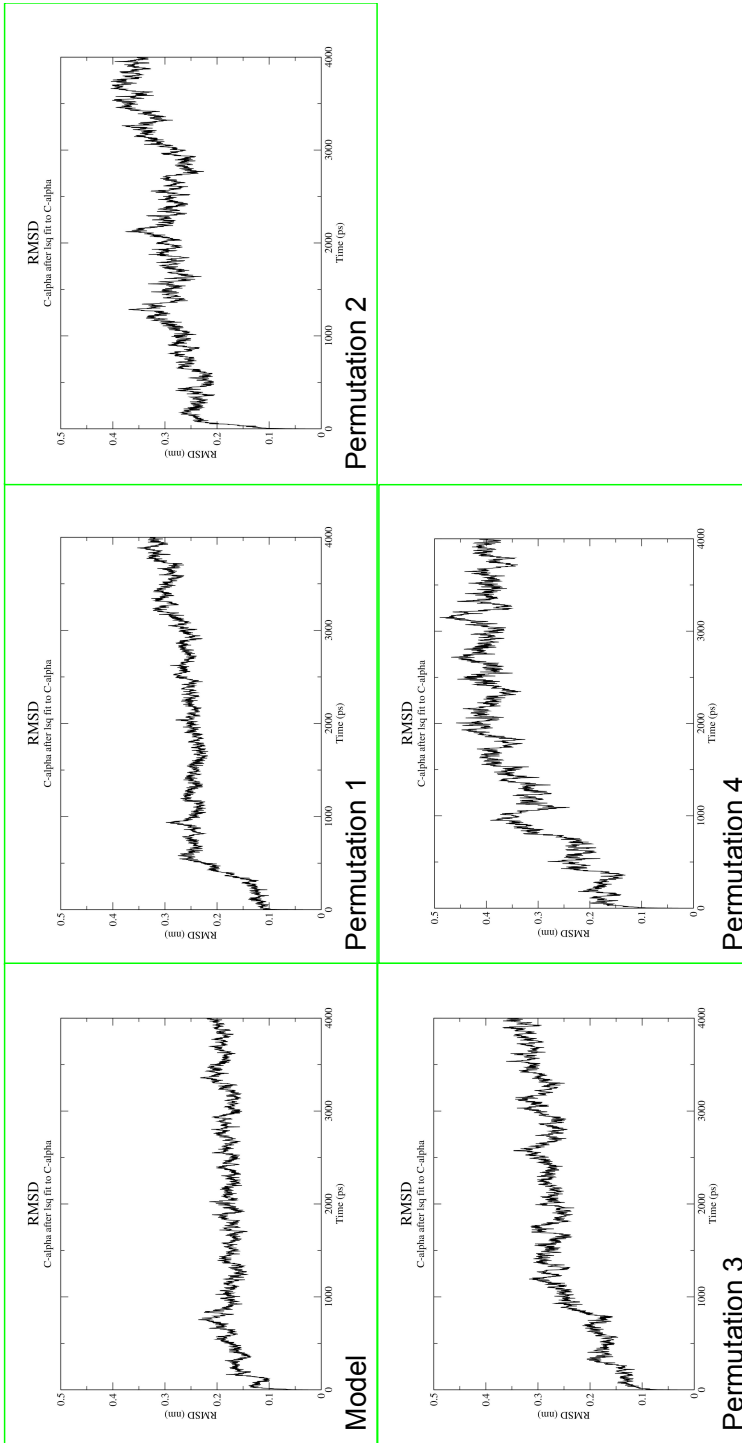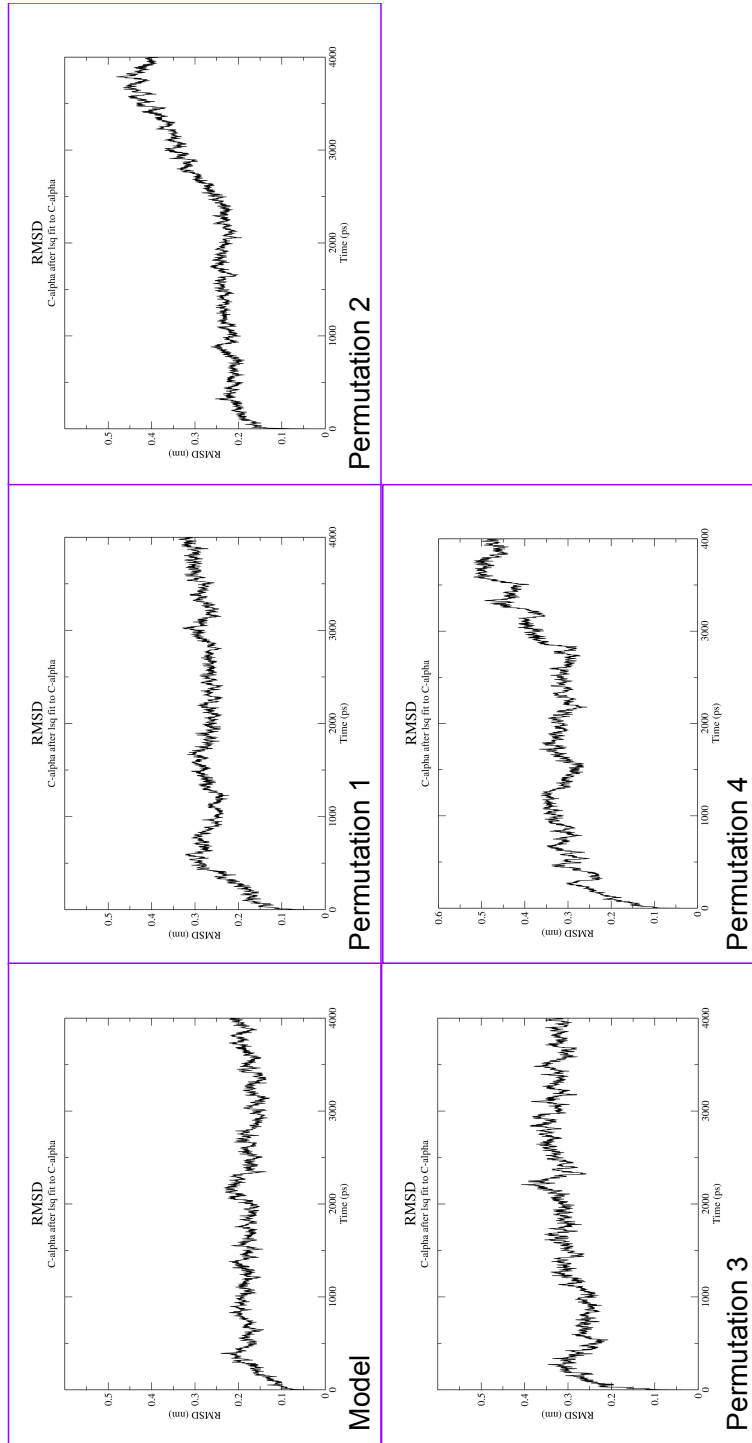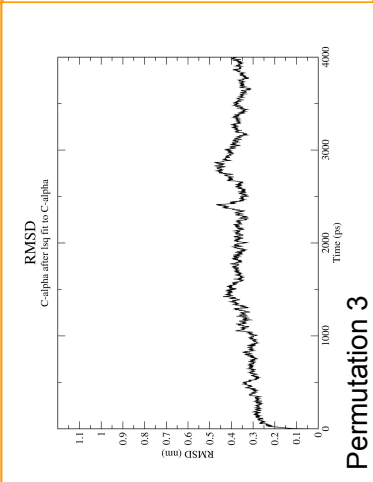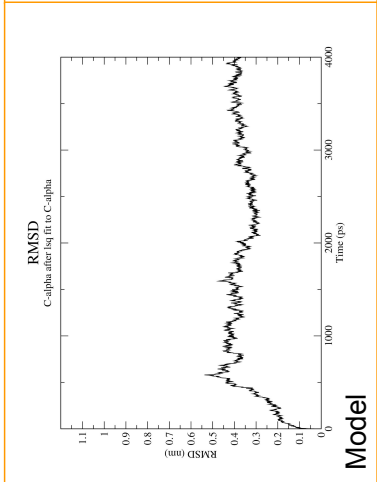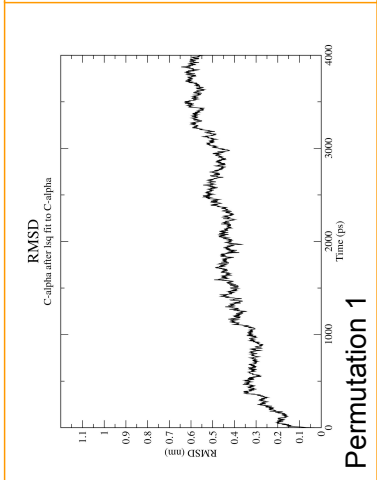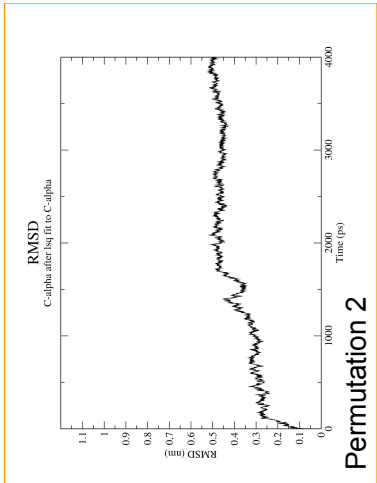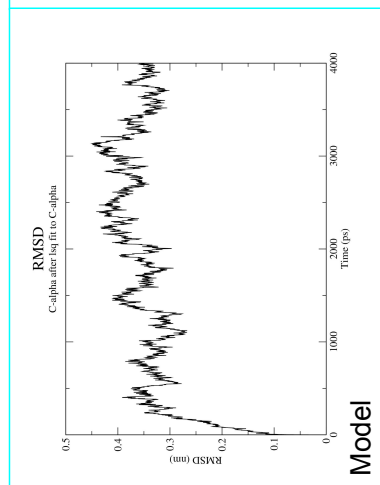
Model

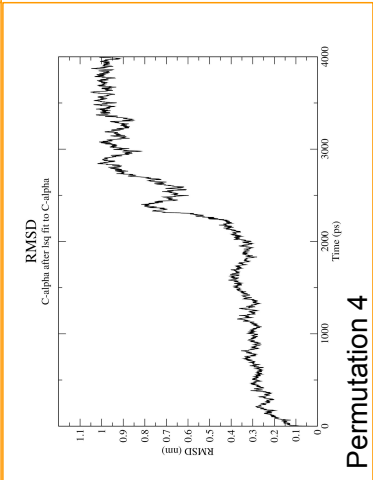Permutation 1

Permutation 2

Permutation 3

Permutation 4

1ARK

Permutation 2

Permutation 1

Permutation 4

Model

Permutation 3

1HX2

159

Model

Permutation 1

Permutation 2

Permutation 3

Permutation 4

1KV0

160

Model

Permutation 1

Permutation 2

Permutation 3

Permutation 4

1TCP

162

Model

Permutation 1

Permutation 2

Permutation 3

Permutation 4

1UOY

163

Model

Permutation 1

Permutation 2

Permutation 3

Permutation 4

1ZFI

Model

Permutation 1

Permutation 2

Permutation 3

Permutation 4

2BHI

Model   Permutation 1   Permutation 2
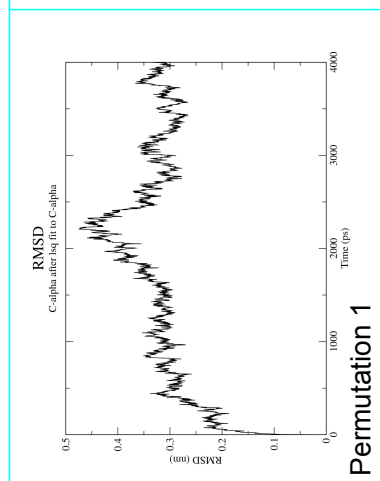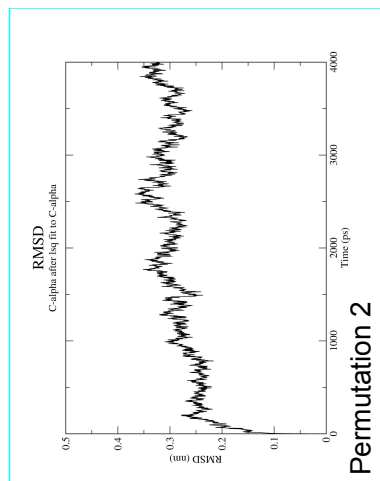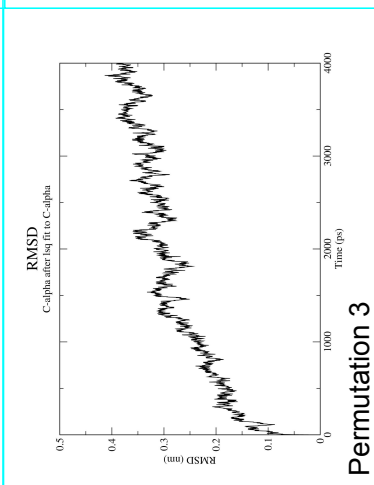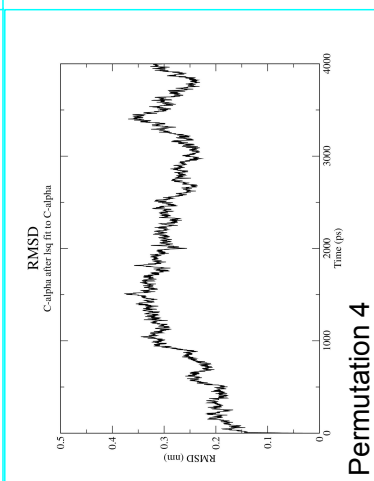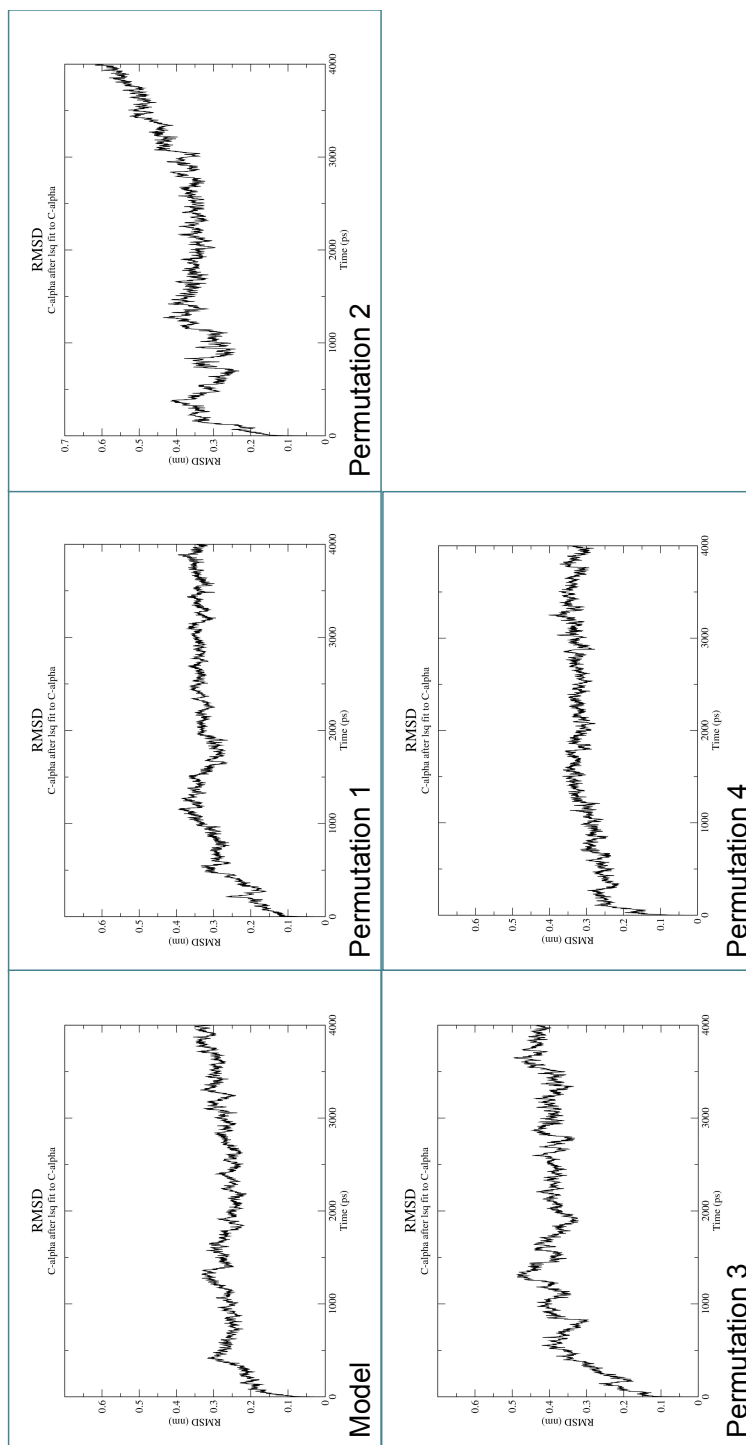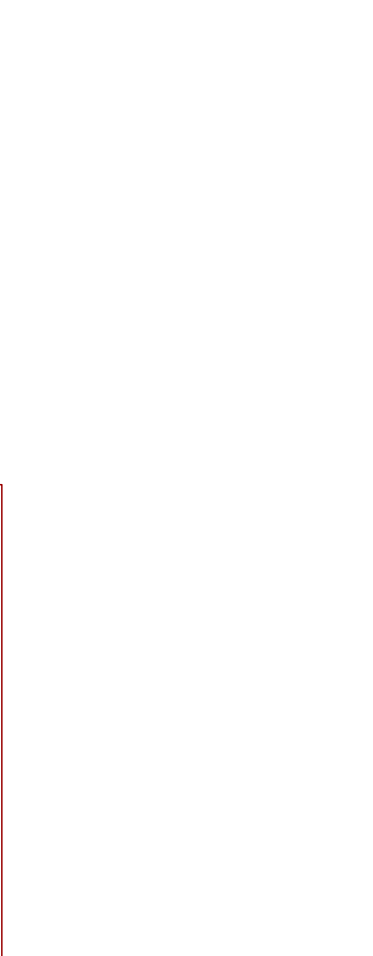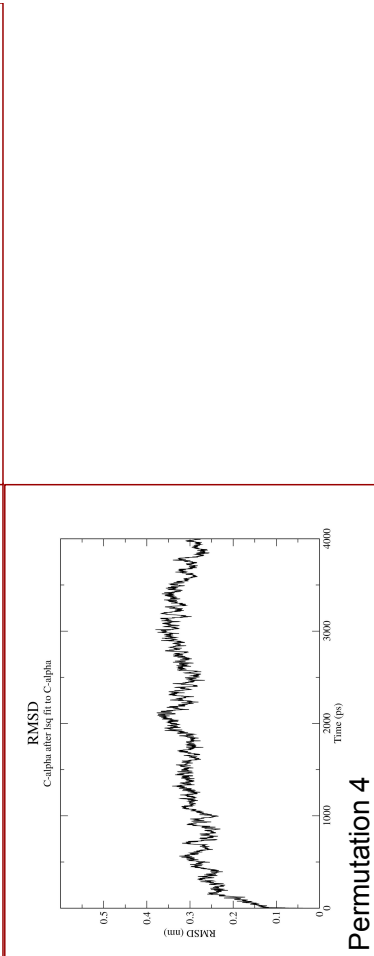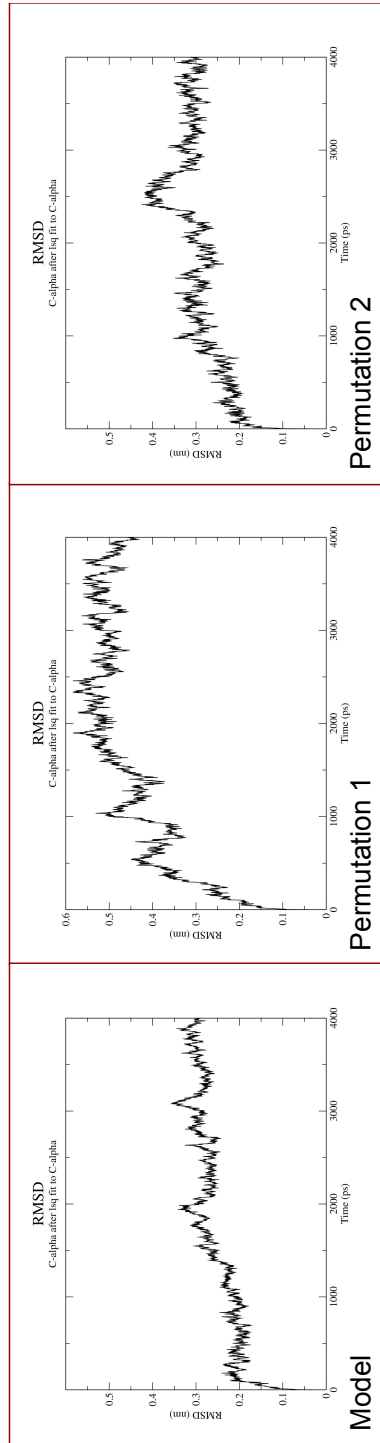
Permutation 3   Permutation 4

2CDX

166
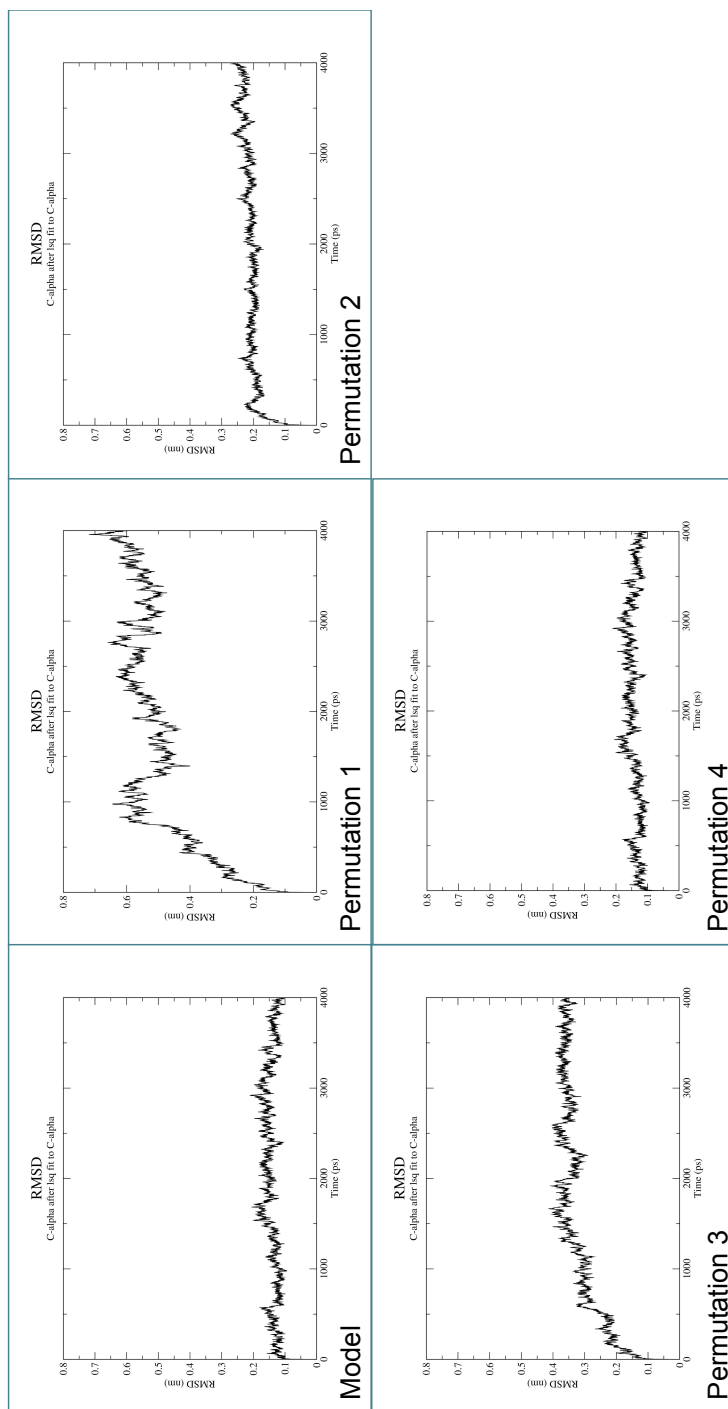
Supplementary File 2: RMSD plots. RMSD variation measured for each protein during 4 ns of molecular dynamics simulation.

167

8.5 RING MD: gathering time into structures. Supplementary Information.

RINs for Molecular Dynamics

A protein r can be represented as a $\mathbb{R}^{n\times 3}$ matrix, where n is the number of amino acids that need a *3*-dimensional representation. Let r_i be the *i*-th residue, then we can compute the so-called *contact map* $m \in \{0,1\}^{n\times n}$, which is defined as follows:

$$m_{ij} = \begin{cases} 1, & if \ \| \ r_i - r_j \ \| < c \\ 0, & otherwise \end{cases}$$

Where $\| \, . \, \|$ is the well known *euclidean norm* and c is a user-defined cut-off (typically set to 6-12Å). Contact maps are intimately related to RINs, and any result obtained in the former can be immediately extended to the latter. In our work, we introduced the concept of Time-Dependent Contact-Map (TDCM), which is based on the following set of contact matrices:

$$m_{ij}(t) = \begin{cases} 1, & if \ \| \ r_i(t) - r_j(t) \ \| < c \\ 0, & otherwise \end{cases}$$

where $r(t)$ is the set of *3*-dimensional coordinates of the protein r at discrete time $t \in T$ of MD simulation. The first straightforward implementation of TDCM is proposed in (169), where authors proposed to calculate the mean of $m(t)$ matrices to easily summarize the MD. In order to generalize this concept, we represent each contact-map $m(t)$ as a vector by means of the *vectorization* operator *vec*():

$$vec(m) = [m_{1,1}, \dots, m_{n,1}, m_{1,2}, \dots, m_{n,2}, \dots m_{1,n}, \dots, m_{n,n}]^T$$

and use the concept of *convex hull* to denote the space of possible contact-maps:

$$conv\left(m(t_1), \dots, m(t_{|T|})\right) = \left\{ \sum_{t \in T} a_t vec(m(t)) \,\middle|\, \forall t : a_t \geq 0 \wedge \sum_{t \in T} a_t = 1 \right\}$$

In other words, the space of conformers observed during the MD is described by a linear combination of the contact-maps. This is an interesting formulation, as it considers explicitly all the possible states of the protein. In particular, if we set $a_t = |T|^{-1} \forall t \in T$, we obtain the formulation proposed in (169). On the other hand, a given protein of interest explores different conformers during the MD, and none of them will be described accurately by the mean contact map proposed in(169). To overcome this limitation, we propose to use clustering methods to better represent TDCM. In our work, we used k-means algorithm (165) to identify *k* similar contact maps:

$$argmin_S \sum_{i=1}^{k} \sum_{m(t_j) \in S_i} \| m(t_j) - \mu_i \|$$

Thus, contact-maps are assigned to different clusters in *S,* and their means μ_1 , ..., μ_k can be used to summarize states-dependent interactions. In other words, the convex hull is basically segmented by assigning a coefficient of $a_j = |S_i|^{-1}$ to all contact-maps that belonging to certain cluster S_j and 0 otherwise. As results, we obtain a set of contact maps that are likely to be generated from similar conformers, and therefore good representative of a that protein state. The number *k* of conformers has been estimated by means of the Bayesian information criterion (BIC) (174), which is a well-known model selection technique. It is defined as follows:

$$-2 \ln(P(m(t)|\mu, S)) + k \cdot n^2 \cdot \ln(|T|)$$

The objective function is dependent on two aspects: the likelihood of the k-means fitting, and a penalty related to the $k \cdot n^2$ number of parameters needed for k clusters. Thus, BIC aims to select a model representing a compromise between fitting quality and its complexity. In order to speed up the computations, we simplified the contact map representation using a map function $SCM: \mathbb{R}^{n \times n} \to \mathbb{R}^n$. This function transform the matrix $m(t)$ into a a vector $m'(t)$ where:

$$m'_i(t) = \begin{cases} 1, & \text{if } \sum_{j=1}^{n} m_{ij}(t) > 0 \\ 0, & \text{otherwise} \end{cases}$$

Indeed, for each residue, we set to 1 those ones that are currently involved in an interaction (see Fig S1).
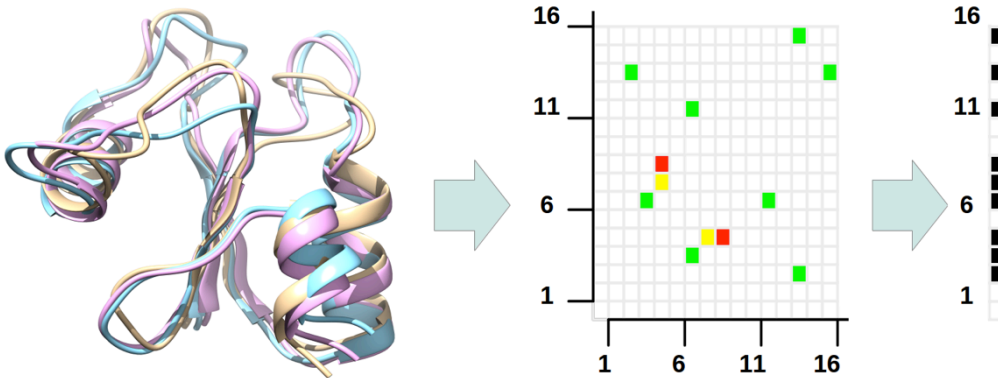
Fig S1: Contact map and simplified contact map. With RING, we can label interactions according to different type of bonds, like hydrogen bonds (green), ionic bonds (yellow) and pi-cations (red). The binary symmetric matrix M representing the contact map is transformed into a binary vector V. Such vector represents the existence of an interaction of the i-th amino-acid. As an example, the bond in $M_{4,7}$ lead to the coloring of V_4 and V_7. Clearly, this lead to a lossy compression of the contact map.

170

This is a critical improvement, since it is practically infeasible to run clustering algorithms on thousands of protein frames where the number of residues is high (e.g. greater than 1,000). In addition, it is common to focus only on contact maps with main chain-side chain and side chain-side chain interactions, as they are the most informative bonds (175). These contact maps are highly sparse matrices, so the information lost by means of our function f is very limited. Finally, this has also an impact in the BIC model selection process, as it reduces the penalty term related to the model complexity and allowing the discovery of additional useful conformers. It should be noted that the parameters obtained from clustering can be interpreted as a mixture model. Therefore, one can use such a model to test if a different MD of the same protein is generated from the same stochastic process. A clear instance of this idea is the evaluation of mutations impact in the protein dynamic or in its stability.

Validation

In order to validate ring MD clustering methodology, we used and idea from CoDNaS (181), a database of known experimental conformations for thousands of proteins. In their work, the authors suggested to use hierarchical clustering on the protein structure coordinates to determine PDB similarities. Such an approach enables the detection of largely different conformers describing diverse native states. Ideally, we expect to observe agreement between ring MD and CoDNaS-based clustering for MD simulated conformers. In agreement to CoDNaS definition, we computed the pairwise distance of MD frame coordinates using the Bio3D package (185), and calculated a dendrogram. Using in-house code, we obtained the conformers dendrogram represented from simplified contact maps (SCM). In Table 1, the clustering similarity is evaluated by means of Cophenetic correlation. For Glutaredoxin, inter-atomic contacts seems the main cause of tridimensional

conformation switch, while the changes observed for hydrogen bonds in both Ubiquitin and Lysozyme better explain the different conformers. Ionic contacts also play a role in two situations. The joint combination of multiple types of interactions (denoted as *all*) has a correlation value of almost 0.5, and demonstrates the high explanatory power of RINs in the detection of different protein native states. A second appealing strength of RINs is their ability to spot the type of interactions promoting conformational changes. This is a clear abstraction of the whole dynamics, and enables the quick detection of key factors for the protein stability and its fluctuations. A last important consideration arises when we compare the average number of interactions for a given residue during the MD with its root mean square fluctuation. As shown in Table 2, the inter-atomic contacts explain a significant part of the amino-acid degree of freedom. This is obvious in practice, and is also well modeled by our RIN-based MD representation.

	Ubiquitin	Lysozyme	Glutaredoxin
H-bond	0.423	0.350	0.213
Van-der-	0.176	0.153	0.604
π-cation	0.061	0.033	0.233
π - π stack	0.046	0.179	0.196
Ionic	0.098	0.323	0.314
All	0.460	0.500	0.474

Table 1: Cophenetic correlation of PDB clustering and RING MD average clustering. In general, there is good agreement between the dendrograms produced from the different type of protein representations. The "All" row represents the sum of all the interactions but the inter-atomic contacts, and describes the existence of any bond.

	Ubiquitin	Lysozyme	Glutaredoxin
H-bond	-0.042	-0.056	-0.064
Van-der-	-0.619	-0.439	-0.294
π-cation	-0.066	-0.031	-0.246
π - π stack	-0.050	-0.013	-0.179
Ionic	-0.058	-0.030	-0.190
All	-0.074	-0.043	-0.282

Table 2: Correlation of RING MD mean number of interactions of a residue with respect to amino acid RMSF. Total amount of Van-der-Waals *contacts correlates with the residue fluctuations.*

Protein analysis

Ubiquitin

Ubiquitin is a protein involved in posttranslational modifications driving intracellular signaling events (186).

Ionic	E1	E2	β1	β2	L1	L2	L3	L4
Cluster 1	K27-(D52) E34-(K11)	R54-(E51)	K11-(E34)	R72-(D39)	Few int.	D39-(R72)	D52-(K27) E51-(R54)	Few int.
Cluster 2	K27-(D52) E34-(K11)	D58-(R54)	K11-(E34)	R72-(D39) R42-(D39)	Few int.	D39-(R72) D39-(R42)	D52-(K27) R54-(D58)	Few int.
Cluster 3	K27-(D52) E34-(K11)	D58-(R54)	K11-(E34)	Few int.	Few int.	Few int.	D52-(K27) R54-(D58)	Few int.
Cluster 4	K27-(D52) E34-(K11)	R54-(E51)	K11-(E34)	Few int.	Few int.	Few int.	D52-(K27) E51-(R54)	Few int.

Table 3: Ionic interactions differences between the four conformational clusters.

K11, as demonstrated in (42), appears to play a key role in Ubiquitin structural stability, and our results confirm this finding, as the residue is involved in interactions in all the clusters (see Table 3). The key residues involved in conformational changes appear to be D39, R42, E51, D58 and R72 [Sup. Fig. 1].
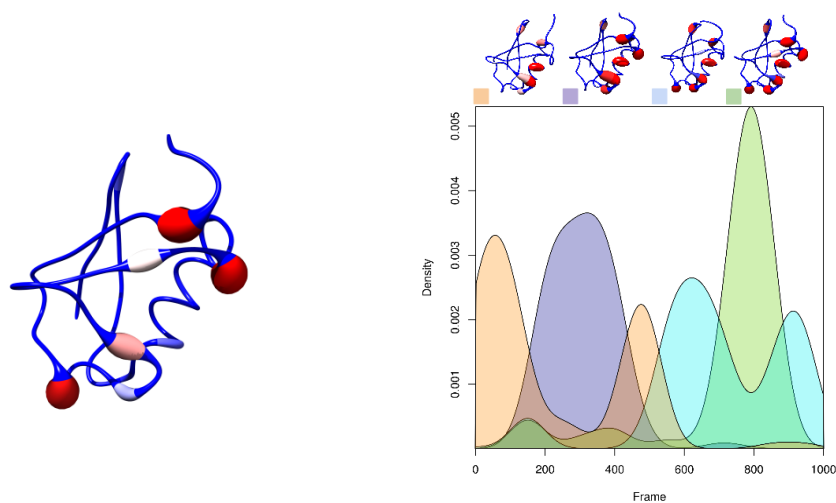
Supplementary Figure 1: Ionic interactions Ubiquitin driven clusterization.

H-bonds	E1	E2	β1	β2	L1	L2	L3	L4
Cluster 1	T22 N25 K27	T55 D58	T7 K11	Q41 S65 L73	Few int.	I36	Few int.	Q62
Cluster 2	T22 N25 K27	T55 D58	T7 K11	Q40 Q41 S65	E18 D21	I36	Few int.	Q62
Cluster 3	N25 K27 K29	T55 S57 D58	T7 G10 K11 E16	Q41 S65	E18	Few int.	Few int.	Q62
Cluster 4	T22 N25 K27	T55 D58	T7 G10 K11	Q41 S65	Few int.	P38	Few int.	Q62

Table 4: H-bonds differences between the four conformational clusters.

The main H bond performing characters for structural maintenance are, according to Table 4, E1, β1 and β2, and E2. E1 N25 and K27 perform constant H bond networks during all the simulation, as well as T55 and D58 for E2. β1 and β2 domains are held by T7 and K11, and Q41 and S65 respectively. Considering loops, L1 performs switching interactions, which are present only in clusters 2 and 3 and are performed by E18 and D21. On

174

the other hand, L4 contains a crucial H bond performing residue, Q62. Residues involved in H-bonds and assuming a key role in conformational transitions are E16, E18, D21, Q40 and L73 [Sup. Fig. 2].
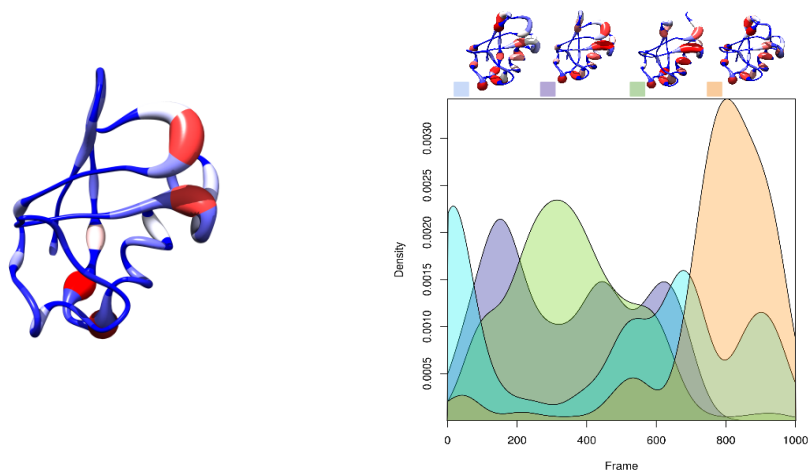
Supplementary Figure 2: H-bond interactions Ubiquitin driven clusterization.

Considering all the computed interaction and all the obtained clusters, the key residues involved in conformational transitions for Ubiquitin appear to be D39, Q40, R42, R72 and L73. These residues appear to be involved in every clustering as driving conformational changes residues [Sup. Fig. 3]. In 2011, Ralat et al. highlighted R72 as one of the second phase cleavage sites for human Insulin-degrading enzyme (IDE) (179), and according to RING MD analysis this residue appears to play also a major structural role, being involved in conformational transitions.
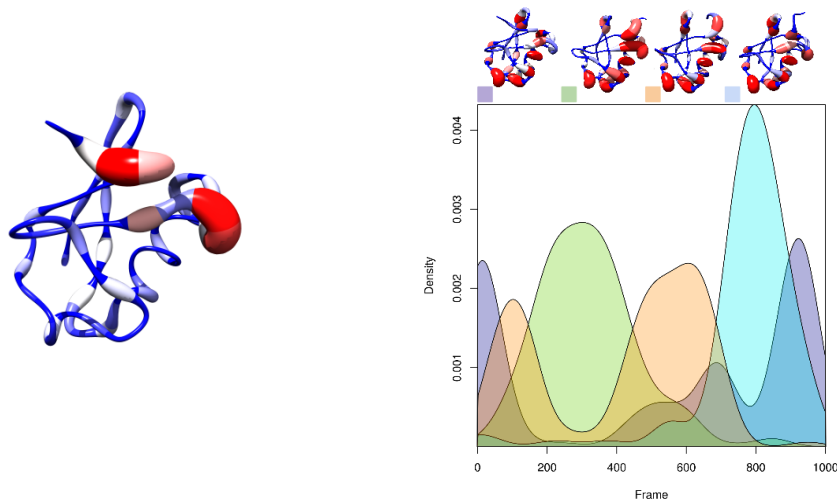
Supplementary Figure 3: All interactions Ubiquitin driven clusterization.

Lysozyme

As mentioned formerly, the second target was T4 Lysozyme. This enzyme is involved in host peptidoglycan degradation and cell lysis (182).

Ionic	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	β1
Cluster 1	E5-R8 D10-(R148) E11-(R14) E11-(R145)	K48-E45 R52-(E62)	E62-(R52) R76-D72 R80-(E108)	Few int.	Few int.	E108-(R80)	Few int.	D127-(R154)	R137-(E22)	R145-(E11) R148-(D10) R154-(D127)	E22-(R137) R14-(E11)
Cluster 2	E5-R8 D10-(R148) E11-(R145)	R52-(E62)	E62-(R52) R76-D72 R80-(E108)	Few int.	D92-R95	E108-(R80)	Few int.	D127-(R154) R125-E128	R137-(E22)	R145-(E11) R148-(D10) R154-(D127)	E22-(R137)
Cluster 3	D10-(R148)	R52-(E62)	E62-(R52) R76-D72 R80-(E108)	K85-D89	D92-R95	E108-(R80)	Few int.	D127-(R154)	Few int.	R148-(D10) R154-(D127)	Few int.
Cluster 4	E5-R8 D10-(R148) E11-(R145)	K48-E45 R52-(E62)	E62-(R52) R76-D72 R80-(E108)	Few int.	D92-R95	E108-(R80)	Few int.	D127-(R154)	R137-(E22)	R145-(E11) R148-(D10) R154-(D127)	E22-(R137)

Table 5: Ionic interactions differences between the four conformational clusters.

176

As shown in Table 5, structural domains apparently being involved in conformational changes are E4, E5, E9 and β1. Helix 10 shows, together with E3, to perform the highest number and more stable interactions. On the contrary, E1 shows a plethora of different interactions scenarios, ranging from one ionic interaction in cluster 3 to four in cluster 1. E11, which shown to perform ionic interactions in clusters 1, 2 and 4, belongs also to the active site of the enzyme (173). Our analysis could thus provide a different role for this residue, highlighting its importance in both structure and function of Lysozyme. Furthermore, E45 is shown to be a mutation spot that whenever mutated provides structural differences (184).
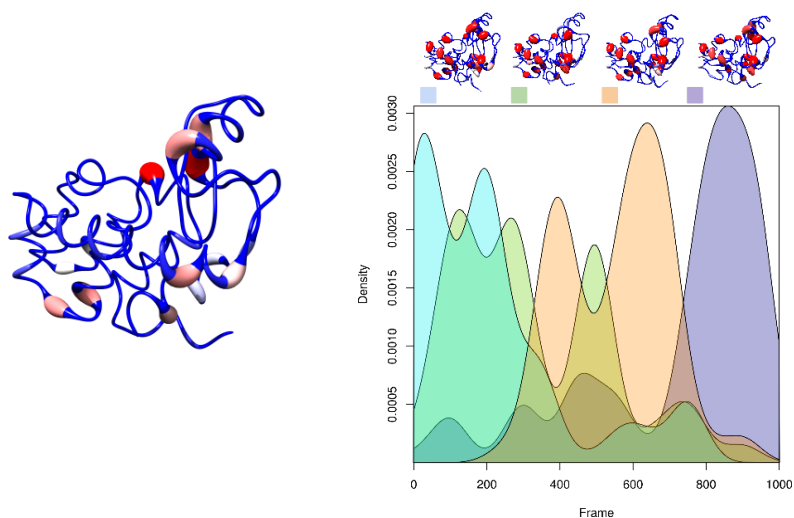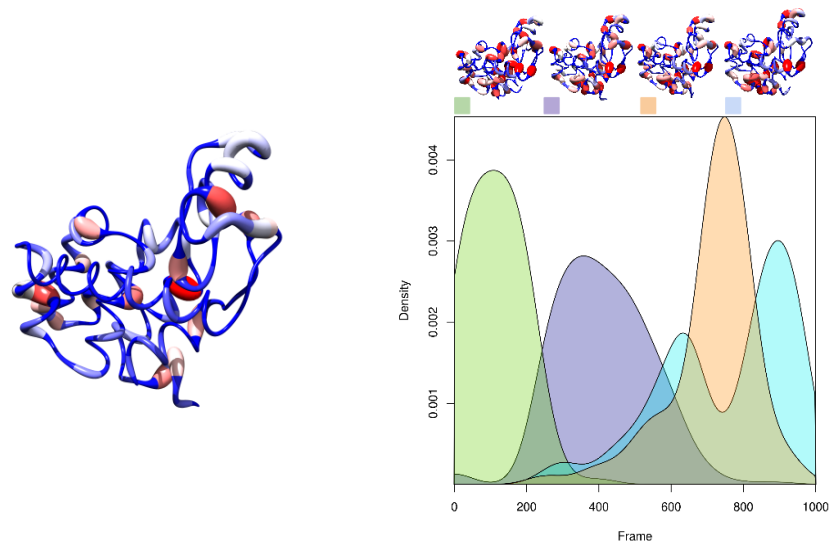

Supplementary Figure 4: Ionic interactions Lysozyme driven clusterization.

H-bonds	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	β1
Cluster 1	N2 E5 D10	S38 A41 D47 R52	T59 E62	N81 K83 L84 S90 L91	A93 R95 R96 C97 Q105	E108 A112	T115 Q122	E128 N132 L133	S136 W128 Y139	T142 T151 T152 F153 R154 T155 T157 D159 Y161	T34 S36 C54 N55
Cluster 2	N2 E5	S38 N40 D47 R52	T59 E62	N81 K83 L84 P86 V87 S90 L91	A93 R95 C97 N101 Q105	E108	K124	N132 L133	S136 W138 Y139	T142 K147 R148 T151 T152 F153 R154 T155 T157 D159 Y161	Y18 G30 T34 S36 C54 N55
Cluster 3	N2 E5	S38 A41	T59 E62	N81 K83 L84 S90 L91	A93 R95 R96 C97 Q105	Few int.	Few int.	E128 N132 L133	S136 W138 Y139	T142 K147 R148 T151 F153 R154 T155 T157 D159 Y161	T34 S36
Cluster 4	N2 E5 D10	P37 S38 N40 S44 D47 R52	T59 E62	N81 K83 L84 P86 S90 L91	A93 R95 R96 C97 N101 Q105	Few int.	T115 S117 Q122	N132 L133	S136 Y139	T142 K147 R148 T151 T152 F153 R154 T155 T157 D159 Y161	Y25 T34 S36 C54

Table 6: H-bonds differences between the four conformational clusters.

Table 6 lists all the residues involved in H bond network formations in the four clusters. Mutations occurring in some of these residues, in particular S44F, S44E, S44W, S117V, and R96A Lysozyme mutants have been

previously demonstrated to provide different enzyme conformations when compared to the WT (184). The H-bond performing residues allowing the conformational switching are D10, Y18, G30, D47, N55, P86, V87, N101, E108, S117, Q122, K124, E128, R145, D159 and Y161 [Sup. Fig. 5].

Supplementary Figure 5: H-bond interactions Lysozyme driven clusterization.

At a more comprehensive level, considering all the interactions, the key residues driving conformational changes appear to be E22, G30, E45, D47, K48, V87, D92, N101, K124, R137 and D159 [Sup. Fig. 6].
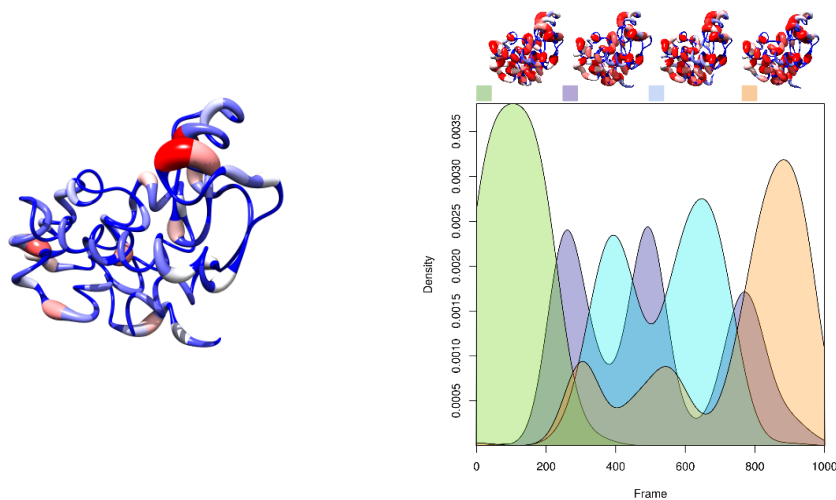
Supplementary Figure 6: All interactions Lysozyme driven clusterization.

Glutaredoxin

The third target of this work was T4 Glutaredoxin, a reducing agent for phage-induced ribonucleotide reductase (183). Like the former two targets, this has been chosen for both its dimensions and for the presence of an active site to investigate on. As for the other two simulations, the first analysis dealt with ionic interactions. In table 7, the peculiar interactions per cluster driving the folding of the secondary structure are highlighted.

Ionic	E1	E2	E3	β1
Cluster 1	D18-K21	D44-K47	Few int.	K3-E32
Cluster 2	D18-K21	E50-(K3)	D80-R83	K3-(E50)
Cluster 3	Few int.	D44-K47 E50-K54	D80-R83	K3-E32
Cluster 4	D18-R22	D44-K47 E50-K54 E50-(K3)	Few int.	K3-(E50)

Table 7: Ionic interactions differences between the four conformational clusters.

K21 and D80 are, interestingly, two residues known to undergo mutations, the former causing loss of ability by the protein to get reduced by thioredoxin reductase and the latter known to provide an opposite effect(183). By the fast RING MD analysis, we were able to identify these residues without carrying out mutation analysis, and just identifying the node centrality during the entire MD simulation. The first residue leading to conformational changes due to ionic interactions was found to be E32, involved in conformational transition between cluster 1-2 and 3-4. In clusters 1 and 3, E32 interacts with K3. On the other hand, in clusters 2 and 4 it performs no interactions. D44 appears to be an ionic switch required for clusters 3 and 4, having K47 as a positive counterpart. E50, belonging to E2, interacts with β1 K3 in clusters 2 and 4 and switches the interaction towards K54 in clusters 3 and 4. Thus cluster 4 seems to be characterized by a transient ionic switch. D80 and R83 are, at the end, the driving force leading to clusters 2 and 3 formation. The main secondary structure elements driving conformational change appear to be, at the end, E3 and β1, with some biases coming from E2 [Sup. Fig. 7].
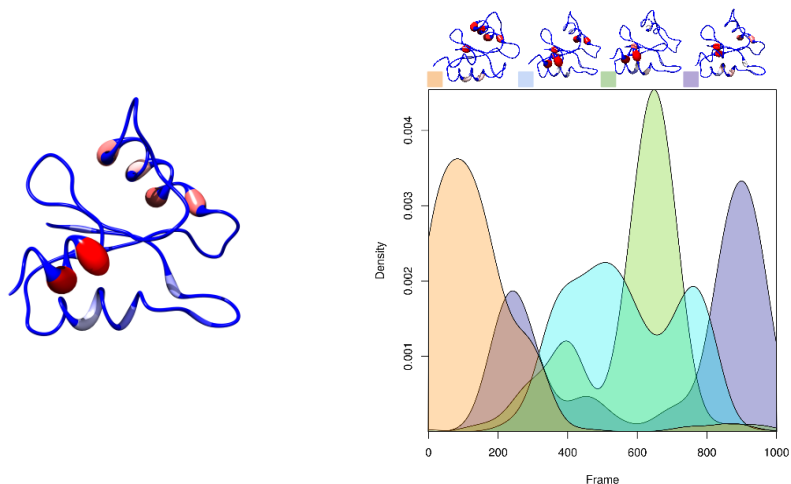
Supplementary Figure 7: Ionic interactions Glutaredoxin driven clusterization.

H-bonds	E1	E2	E3	β1	L1	L2	L3	L4	L5
Cluster 1	K21 T25	A49 E50 T53	Few int.	Y5	S9 C14	Few int.	Few int.	Few int.	Few int.
Cluster 2	K21 T25	A49 T53	Few int.	Few int.	D8 N10 C14	Few int.	Few int.	Few int.	Few int.
Cluster 3	K21 T25	A49 E50 T53	Few int.	Y5	S9 N10	Few int.	N35	Few int.	D72 S74
Cluster 4	K21 T25	A49 T53	Q81	G78	D8 N10 C14	Few int.	N35	Few int.	Few int.

Table 8: H-bonds differences between the four conformational clusters.

As noticeable by table 8, two key residues involved in H bond formation are K21 and T25, which are crucial for E1. In E2, A49 and T53 are always present, and they appear to have a structural role for helix maintenance. Loops, as one could expect, are not performing H bonds frequently, except for L5 in cluster 4 and L3 in clusters 3 and 4. The overall H-bond driven conformational changes are switched by Y5, E50, D72, S74, G78 and Q81 [Sup. Fig. 8].
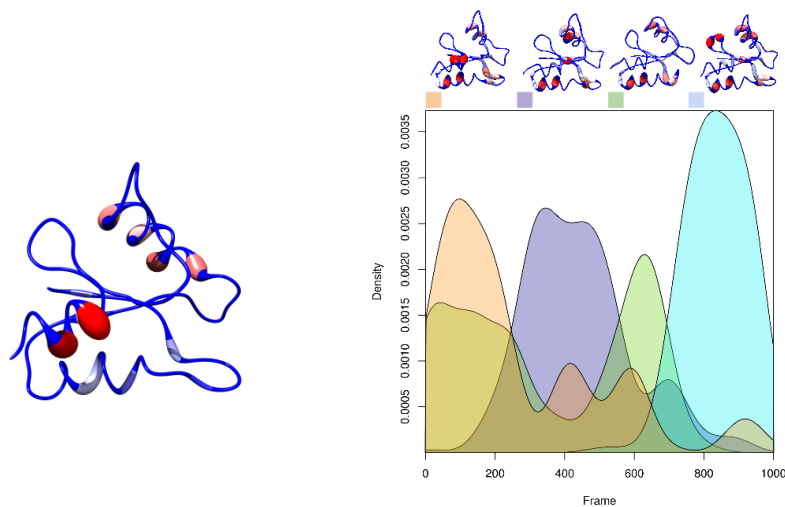
Supplementary Figure 8: H-bond interactions Glutaredoxin driven clusterization.

Coming to the final analysis, a comprehensive evaluation of the whole set of the considered interaction lead to identify Y5, D18, R22, E32, N35, F43, T64, Q67, D72, S74, D80 and Y85 as key residues for overall conformational transitions during the entire 50 ns MD simulation [Sup. Fig. 9]. Moreover, as previously mentioned, the functional importance of K21 added to the structural role highlighted by RING MD analysis suggests the central role of this residue for this protein.
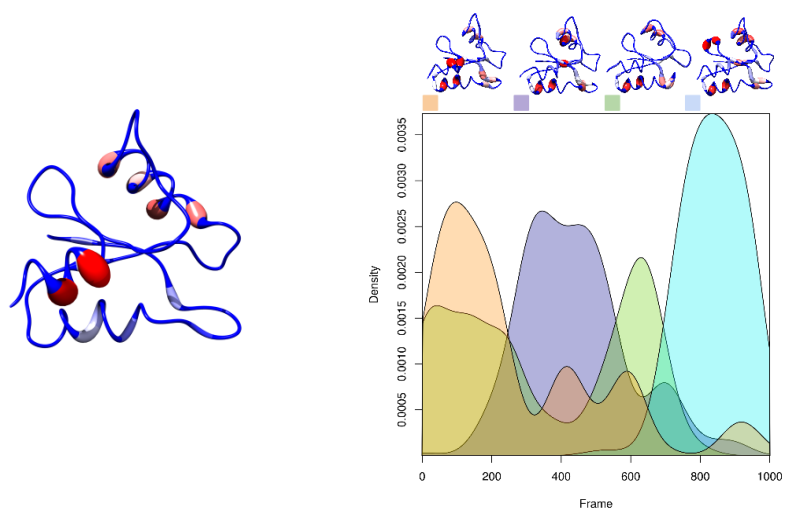
Supplementary Figure 9: All interactions Glutaredoxin driven clusterization.

Bibliography

1.   John Fitzgerald Kennedy. Address in New York City to the National Associacion of Manufacturers (496), Public Papers of the Presidents: John F. Kennedy, 1961. 1961.

2.   Soga T. Cancer metabolism: key players in metabolic reprogramming. Cancer Sci. 2013 Mar;104(3):275–81.

3.   Gnarra JR, Tory K, Weng Y, Schmidt L, Wei MH, Li H, et al. Mutations of the VHL tumour suppressor gene in renal carcinoma. Nat Genet. 1994 May;7(1):85–90.

4.   Stolle C, Glenn G, Zbar B, Humphrey JS, Choyke P, Walther M, et al. Improved detection of germline mutations in the von Hippel-Lindau disease tumor suppressor gene. Hum Mutat. 1998;12(6):417–23.

5.   Frew IJ, Krek W. pVHL: a multipurpose adaptor protein. Sci Signal. 2008;1(24):pe30.

6.   Fedulova N, Hanrieder J, Bergquist J, Emrén LO. Expression and purification of catalytically active human PHD3 in Escherichia coli. Protein Expr Purif. 2007 Jul;54(1):1–10.

7.   Chowdhury R, McDonough MA, Mecinović J, Loenarz C, Flashman E, Hewitson KS, et al. Structural basis for binding of hypoxia-inducible factor to the oxygen-sensing prolyl hydroxylases. Struct Lond Engl 1993. 2009 Jul 15;17(7):981–9.

8.   Zilfou JT, Lowe SW. Tumor suppressive functions of p53. Cold Spring Harb Perspect Biol. 2009 Nov;1(5):a001883.

9.   Zerrouqi A, Pyrzynska B, Febbraio M, Brat DJ, Van Meir EG. P14ARF inhibits human glioblastoma-induced angiogenesis by upregulating the expression of TIMP3. J Clin Invest. 2012 Apr 2;122(4):1283–95.

10.   Lomazzi M, Moroni MC, Jensen MR, Frittoli E, Helin K. Suppression of the p53- or pRB-mediated G1 checkpoint is required for E2F-induced S-phase entry. Nat Genet. 2002 Jun;31(2):190–4.

11.	Lai Y, Song M, Hakala K, Weintraub ST, Shiio Y. Proteomic dissection of the von Hippel-Lindau (VHL) interactome. J Proteome Res. 2011 Nov 4;10(11):5175–82.

12.	Latif F, Tory K, Gnarra J, Yao M, Duh FM, Orcutt ML, et al. Identification of the von Hippel-Lindau disease tumor suppressor gene. Science. 1993 May 28;260(5112):1317–20.

13.	Semenza GL. Regulation of mammalian O2 homeostasis by hypoxia-inducible factor 1. Annu Rev Cell Dev Biol. 1999;15:551–78.

14.	Maher ER, Neumann HP, Richard S. von Hippel-Lindau disease: a clinical and scientific review. Eur J Hum Genet EJHG. 2011 Jun;19(6):617–23.

15.	Kim WY, Kaelin WG. Role of VHL gene mutation in human cancer. J Clin Oncol Off J Am Soc Clin Oncol. 2004 Dec 15;22(24):4991–5004.

16.	Minervini G, Panizzoni E, Giollo M, Masiero A, Ferrari C, Tosatto SCE. Design and Analysis of a Petri Net Model of the Von Hippel-Lindau (VHL) Tumor Suppressor Interaction Network. PloS One. 2014;9(6):e96986.

17.	Knudson AG. Mutation and cancer: statistical study of retinoblastoma. Proc Natl Acad Sci U S A. 1971 Apr;68(4):820–3.

18.	Gnarra JR, Ward JM, Porter FD, Wagner JR, Devor DE, Grinberg A, et al. Defective placental vasculogenesis causes embryonic lethality in VHL-deficient mice. Proc Natl Acad Sci U S A. 1997 Aug 19;94(17):9102–7.

19.	Van der Horst MA, Hellingwerf KJ. Photoreceptor proteins, "star actors of modern times": a review of the functional dynamics in the structure of representative members of six different photoreceptor families. Acc Chem Res. 2004 Jan;37(1):13–20.

20.	Quail PH. The phytochrome family: dissection of functional roles and signalling pathways among family members. Philos Trans R Soc Lond B Biol Sci. 1998 Sep 29;353(1374):1399–403.

21.	Hoff WD, Xie A, Van Stokkum IH, Tang XJ, Gural J, Kroon AR, et al. Global conformational changes upon receptor stimulation in photoactive yellow protein.

Biochemistry (Mosc). 1999 Jan 19;38(3):1009–17.

22.   Kort R, Hoff WD, Van West M, Kroon AR, Hoffer SM, Vlieg KH, et al. The xanthopsins: a new family of eubacterial blue-light photoreceptors. EMBO J. 1996 Jul 1;15(13):3209–18.

23.   Ahmad M, Cashmore AR. HY4 gene of A. thaliana encodes a protein with characteristics of a blue-light photoreceptor. Nature. 1993 Nov 11;366(6451):162–6.

24.   Huala E, Oeller PW, Liscum E, Han IS, Larsen E, Briggs WR. Arabidopsis NPH1: a protein kinase with a putative redox-sensing domain. Science. 1997 Dec 19;278(5346):2120–3.

25.   Gomelsky M, Klug G. BLUF: a novel FAD-binding domain involved in sensory transduction in microorganisms. Trends Biochem Sci. 2002 Oct;27(10):497–500.

26.   Zoltowski BD, Vaidya AT, Top D, Widom J, Young MW, Crane BR. Structure of full-length Drosophila cryptochrome. Nature. 2011 Dec 15;480(7377):396–9.

27.   Levy C, Zoltowski BD, Jones AR, Vaidya AT, Top D, Widom J, et al. Updated structure of Drosophila cryptochrome. Nature. 2013 Mar 21;495(7441):E3–4.

28.   Glas AF, Schneider S, Maul MJ, Hennecke U, Carell T. Crystal Structure of the T(6-4)C Lesion in Complex with a (6-4) DNA Photolyase and Repair of UV-Induced (6-4) and Dewar Photolesions. Chem - Eur J. 2009 Oct 12;15(40):10387–96.

29.   Hemsley MJ, Mazzotta GM, Mason M, Dissel S, Toppo S, Pagano MA, et al. Linear motifs in the C-terminus of D. melanogaster cryptochrome. Biochem Biophys Res Commun. 2007 Apr;355(2):531–7.

30.   Vaidya AT, Top D, Manahan CC, Tokuda JM, Zhang S, Pollack L, et al. Flavin reduction activates Drosophila cryptochrome. Proc Natl Acad Sci. 2013 Dec 2;201313336.

31.   Mazzotta G, Rossi A, Leonardi E, Mason M, Bertolucci C, Caccin L, et al. Fly cryptochrome and the visual system. Proc Natl Acad Sci. 2013 Apr 9;110(15):6163–8.

32.   Tamanini F, Chaves I, Bajek MI, van der Horst GTJ. Structure function

analysis of mammalian cryptochromes. Cold Spring Harb Symp Quant Biol. 2007;72:133–9.

33.   Peschel N, Helfrich-Förster C. Setting the clock--by nature: circadian rhythm in the fruitfly Drosophila melanogaster. FEBS Lett. 2011 May 20;585(10):1435–42.

34.   Koh K, Zheng X, Sehgal A. JETLAG resets the Drosophila circadian clock by promoting light-induced degradation of TIMELESS. Science. 2006 Jun 23;312(5781):1809–12.

35.   Peschel N, Veleri S, Stanewsky R. Veela defines a molecular link between Cryptochrome and Timeless in the light-input pathway to Drosophila's circadian clock. Proc Natl Acad Sci U S A. 2006 Nov 14;103(46):17313–8.

36.   Emery P, So WV, Kaneko M, Hall JC, Rosbash M. CRY, a Drosophila clock and light-regulated cryptochrome, is a major contributor to circadian rhythm resetting and photosensitivity. Cell. 1998 Nov 25;95(5):669–79.

37.   De Lucrezia D, Slanzi D, Poli I, Polticelli F, Minervini G. Do natural proteins differ from random sequences polypeptides? Natural vs. random proteins classification using an evolutionary neural network. PloS One. 2012;7(5):e36634.

38.   MacKerell, Bashford D, Bellott, Dunbrack, Evanseck JD, Field MJ, et al. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins†. J Phys Chem B. 1998 Apr 1;102(18):3586–616.

39.   Berman H, Henrick K, Nakamura H, Markley JL. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. Nucleic Acids Res. 2007 Jan;35(Database issue):D301–3.

40.   Leach AR. Molecular Modelling: Principles and Applications. 2nd ed. Great Britain: Pearson; 2001. 744 p.

41.   Alder BJ, Wainwright TE. Phase Transition for a Hard Sphere System. J Chem Phys. 1957 Nov 1;27(5):1208–9.

42.   Mandal M, Mukhopadhyay C. Microsecond molecular dynamics simulation of guanidinium chloride induced unfolding of ubiquitin. Phys Chem Chem Phys PCCP. 2014 Oct 21;16(39):21706–16.

43.   Committee on Proposal Evaluation for Allocation of Supercomputing Time for the Study of Molecular Dynamics, Third Round, Board on Life Sciences, Division on Earth and Life Studies, National Research Council. Report of the Committee on Proposal Evaluation for Allocation of Supercomputing Time for the Study of Molecular Dynamics: Third Round [Internet]. Washington (DC): National Academies Press (US); 2012 [cited 2015 Jan 21]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK253909/

44.   Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. Comparison of simple potential functions for simulating liquid water. J Chem Phys. 1983 Jul 15;79(2):926–35.

45.   Neria E, Fischer S, Karplus M. Simulation of activation free energies in molecular systems. J Chem Phys. 1996 Aug 1;105(5):1902–21.

46.   Hess B, Kutzner C, van der Spoel D, Lindahl E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. J Chem Theory Comput. 2008;4(3):435–47.

47.   Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, et al. Scalable molecular dynamics with NAMD. J Comput Chem. 2005;26(16):1781–802.

48.   Proulx SR, Promislow DEL, Phillips PC. Network thinking in ecology and evolution. Trends Ecol Evol. 2005 Jun;20(6):345–53.

49.   Chaouiya C. Petri net modelling of biological networks. Brief Bioinform. 2007 Jul;8(4):210–9.

50.   Petri CA. Communication with automata. Schriften des IIM Nr 3, Bonn: Institut für instrumentelle Mathematik 1962.

51.   Grunwald S, Speer A, Ackermann J, Koch I. Petri net modelling of gene regulation of the Duchenne muscular dystrophy. Biosystems. 2008 May;92(2):189–205.

52.   Martin AJM, Vidotto M, Boscariol F, Di Domenico T, Walsh I, Tosatto SCE. RING: networking interacting residues, evolutionary information and energetics in protein structures. Bioinforma Oxf Engl. 2011 Jul 15;27(14):2003–5.

53.   Sackmann A, Heiner M, Koch I. Application of Petri net based analysis techniques to signal transduction pathways. BMC Bioinformatics. 2006;7:482.

54.   Minervini G, Masiero A, Moro S, Tosatto SCE. In silico investigation of PHD-3 specific HIF1-α proline 567 hydroxylation: A new player in the VHL/HIF-1α interaction pathway? FEBS Lett. 2013 Sep 17;587(18):2996–3001.

55.   Lonser RR, Glenn GM, Walther M, Chew EY, Libutti SK, Linehan WM, et al. von Hippel-Lindau disease. Lancet. 2003 Jun 14;361(9374):2059–67.

56.   Vortmeyer AO, Huang SC, Pack SD, Koch CA, Lubensky IA, Oldfield EH, et al. Somatic point mutation of the wild-type allele detected in tumors of patients with VHL germline deletion. Oncogene. 2002 Feb 14;21(8):1167–70.

57.   Nyhan MJ, O'Sullivan GC, McKenna SL. Role of the VHL (von Hippel–Lindau) gene in renal cancer: a multifunctional tumour suppressor. Biochem Soc Trans. 2008 Jun 1;36(3):472.

58.   Feldman DE, Thulasiraman V, Ferreyra RG, Frydman J. Formation of the VHL-elongin BC tumor suppressor complex is mediated by the chaperonin TRiC. Mol Cell. 1999 Dec;4(6):1051–61.

59.   Melville MW, McClellan AJ, Meyer AS, Darveau A, Frydman J. The Hsp70 and TRiC/CCT chaperone systems cooperate in vivo to assemble the von Hippel-Lindau tumor suppressor complex. Mol Cell Biol. 2003 May;23(9):3141–51.

60.   Thoma CR, Toso A, Gutbrodt KL, Reggi SP, Frew IJ, Schraml P, et al. VHL loss causes spindle misorientation and chromosome instability. Nat Cell Biol. 2009 Aug;11(8):994–1001.

61.   Leonardi E, Murgia A, Tosatto SCE. Adding structural information to the von Hippel-Lindau (VHL) tumor suppressor interaction network. FEBS Lett. 2009 Nov 19;583(22):3704–10.

62.   Leonardi E, Martella M, Tosatto SCE, Murgia A. Identification and in silico analysis of novel von Hippel-Lindau (VHL) gene variants from a large population. Ann Hum Genet. 2011 Jul;75(4):483–96.

63.   Forman JR, Worth CL, Bickerton GRJ, Eisen TG, Blundell TL. Structural

bioinformatics mutation analysis reveals genotype-phenotype correlations in von Hippel-Lindau disease and suggests molecular mechanisms of tumorigenesis. Proteins. 2009 Oct;77(1):84–96.

64.   Limaverde-Sousa G, de Andrade Barreto E, Ferreira CG, Cláudio Casali-da-Rocha J. Simulation of the mutation F76del on the von Hippel–Lindau tumor suppressor protein: Mechanism of the disease and implications for drug development. Proteins Struct Funct Bioinforma. 2013;81(2):349–63.

65.   Domene C, Illingworth CJR. Effects of point mutations in pVHL on the binding of HIF-1α. Proteins. 2012 Mar;80(3):733–46.

66.   Liu J, Nussinov R. Allosteric effects in the marginally stable von Hippel–Lindau tumor suppressor protein and allostery-based rescue mutant design. Proc Natl Acad Sci. 2008 Jan 22;105(3):901–6.

67.   Ivan M, Kondo K, Yang H, Kim W, Valiando J, Ohh M, et al. HIFα Targeted for VHL-Mediated Destruction by Proline Hydroxylation: Implications for O2 Sensing. Science. 2001 Apr 20;292(5516):464–8.

68.   Epstein ACR, Gleadle JM, McNeill LA, Hewitson KS, O'Rourke J, Mole DR, et al. C. elegans EGL-9 and Mammalian Homologs Define a Family of Dioxygenases that Regulate HIF by Prolyl Hydroxylation. Cell. 2001 May 10;107(1):43–54.

69.   Min J-H, Yang H, Ivan M, Gertler F, Kaelin WG, Pavletich NP. Structure of an HIF-1α-pVHL Complex: Hydroxyproline Recognition in Signaling. Science. 2002 Jun 7;296(5574):1886–9.

70.   McNeill LA, Hewitson KS, Gleadle JM, Horsfall LE, Oldham NJ, Maxwell PH, et al. The use of dioxygen by HIF prolyl hydroxylase (PHD1). Bioorg Med Chem Lett. 2002 Jun 17;12(12):1547–50.

71.   Taylor SR, Singh J, Sagoo MS, Lightman SL. Clinical and Molecular Features Associated with Cystic Visceral Lesions in Von Hippel-Lindau Disease. Open Ophthalmol J. 2012 Sep 7;6:83–5.

72.   Hirsilä M, Koivunen P, Günzler V, Kivirikko KI, Myllyharju J. Characterization of the Human Prolyl 4-Hydroxylases That Modify the Hypoxia-inducible Factor. J

Biol Chem. 2003 Aug 15;278(33):30772–80.

73.   Pescador N, Cuevas Y, Naranjo S, Alcaide M, Villar D, Landázuri MO, et al. Identification of a functional hypoxia-responsive element that regulates the expression of the egl nine homologue 3 (egln3/phd3) gene. Biochem J. 2005 Aug 15;390(Pt 1):189–97.

74.   Park S, Radmer RJ, Klein TE, Pande VS. A new set of molecular mechanics parameters for hydroxyproline and its use in molecular dynamics simulations of collagen-like peptides. J Comput Chem. 2005 Nov 30;26(15):1612–6.

75.   Chang S-W, Flynn BP, Ruberti JW, Buehler MJ. Molecular mechanism of force induced stabilization of collagen against enzymatic breakdown. Biomaterials. 2012 May;33(15):3852–9.

76.   in't Veld PJ, Stevens MJ. Simulation of the Mechanical Strength of a Single Collagen Molecule. Biophys J. 2008 Jan 7;95(1):33–9.

77.   Laskowski RA, MacArthur MW, Moss DS, Thornton JM. PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Crystallogr. 1993 Apr 1;26(2):283–91.

78.   Tosatto SC, Battistutta R. TAP score: torsion angle propensity normalization applied to local protein structure evaluation. BMC Bioinformatics. 2007 May 15;8(1):155.

79.   Baker NA, Sept D, Joseph S, Holst MJ, McCammon JA. Electrostatics of nanosystems: Application to microtubules and the ribosome. Proc Natl Acad Sci. 2001 Aug 28;98(18):10037–41.

80.   Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, et al. The Pfam protein families database. Nucleic Acids Res. 2012 Jan;40(Database issue):D290–301.

81.   Crooks GE, Hon G, Chandonia J-M, Brenner SE. WebLogo: a sequence logo generator. Genome Res. 2004 Jun;14(6):1188–90.

82.   Walsh I, Martin AJM, Di Domenico T, Vullo A, Pollastri G, Tosatto SCE. CSpritz: accurate prediction of protein disorder segments with annotation for

homology, secondary structure and linear motifs. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W190–6.

83.   Davey NE, Van Roey K, Weatheritt RJ, Toedt G, Uyar B, Altenberg B, et al. Attributes of short linear motifs. Mol Biosyst. 2012 Jan;8(1):268–81.

84.   Hara S, Hamada J, Kobayashi C, Kondo Y, Imura N. Expression and Characterization of Hypoxia-Inducible Factor (HIF)-3α in Human Kidney: Suppression of HIF-Mediated Gene Expression by HIF-3α. Biochem Biophys Res Commun. 2001 Oct 5;287(4):808–13.

85.   Hudnell HK, Dortch Q, Zenick H. An Overview of the Interagency, International Symposium on Cyanobacterial Harmful Algal Blooms (ISOC-HAB): Advancing the Scientific Understanding of Freshwater Harmful Algal Blooms. In: Hudnell HK, editor. Cyanobacterial Harmful Algal Blooms: State of the Science and Research Needs [Internet]. Springer New York; 2008 [cited 2015 Jan 3]. p. 1–16. Available from: http://link.springer.com/chapter/10.1007/978-0-387-75865-7_1

86.   Zambonino-Infante JL, Claireaux G, Ernande B, Jolivet A, Quazuguel P, Sévère A, et al. Hypoxia tolerance of common sole juveniles depends on dietary regime and temperature at the larval stage: evidence for environmental conditioning. Proc R Soc Lond B Biol Sci. 2013 May 7;280(1758):20123022.

87.   Rantanen K, Pursiheimo J, Högel H, Himanen V, Metzen E, Jaakkola PM. Prolyl Hydroxylase PHD3 Activates Oxygen-dependent Protein Aggregation. Mol Biol Cell. 2008 May 1;19(5):2231–40.

88.   Cioffi CL, Qin Liu X, Kosinski PA, Garay M, Bowen BR. Differential regulation of HIF-1α prolyl-4-hydroxylase genes by hypoxia in human cardiovascular cells. Biochem Biophys Res Commun. 2003 Apr 11;303(3):947–53.

89.   Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999 Oct 15;286(5439):531–7.

90.   Thomas JG, Olson JM, Tapscott SJ, Zhao LP. An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic

expression profiles. Genome Res. 2001 Jul;11(7):1227–36.

91.   Wu M, Liu L, Hijazi H, Chan C. A multi-layer inference approach to reconstruct condition-specific genes and their regulation. Bioinformatics. 2013 Jun 15;29(12):1541–52.

92.   Almaas E. Biological impacts and context of network theory. J Exp Biol. 2007 May;210(Pt 9):1548–58.

93.   Weitz JS, Benfey PN, Wingreen NS. Evolution, interactions, and biological networks. PLoS Biol. 2007 Jan;5(1):e11.

94.   Mahon PC, Hirota K, Semenza GL. FIH-1: a novel protein that interacts with HIF-1alpha and VHL to mediate repression of HIF-1 transcriptional activity. Genes Dev. 2001 Oct 15;15(20):2675–86.

95.   Heiner M, Sriram K. Structural analysis to determine the core of hypoxia response network. PloS One. 2010;5(1):e8600.

96.   Kohn KW, Riss J, Aprelikova O, Weinstein JN, Pommier Y, Barrett JC. Properties of switch-like bioregulatory networks studied by simulation of the hypoxia response control system. Mol Biol Cell. 2004 Jul;15(7):3042–52.

97.   Yu Y, Wang G, Simha R, Peng W, Turano F, Zeng C. Pathway switching explains the sharp response characteristic of hypoxia response network. PLoS Comput Biol. 2007 Aug;3(8):e171.

98.   Schoenfeld AR, Davidowitz EJ, Burk RD. Elongin BC complex prevents degradation of von Hippel-Lindau tumor suppressor gene products. Proc Natl Acad Sci U S A. 2000 Jul 18;97(15):8507–12.

99.   Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, Chen C, et al. The IntAct molecular interaction database in 2012. Nucleic Acids Res. 2012 Jan;40(Database issue):D841–6.

100.   Frew IJ, Krek W. Multitasking by pVHL in tumour suppression. Curr Opin Cell Biol. 2007 Dec;19(6):685–90.

101.   Tang N, Mack F, Haase VH, Simon MC, Johnson RS. pVHL Function Is Essential for Endothelial Extracellular Matrix Deposition. Mol Cell Biol. 2006 Apr

1;26(7):2519–30.

102.	Rohr C, Marwan W, Heiner M. Snoopy--a unifying Petri net framework to investigate biomolecular networks. Bioinforma Oxf Engl. 2010 Apr 1;26(7):974–5.

103.	Maynard MA, Evans AJ, Hosomi T, Hara S, Jewett MAS, Ohh M. Human HIF-3alpha4 is a dominant-negative regulator of HIF-1 and is down-regulated in renal cell carcinoma. FASEB J Off Publ Fed Am Soc Exp Biol. 2005 Sep;19(11):1396–406.

104.	Li QF, Wang XR, Yang YW, Lin H. Hypoxia upregulates hypoxia inducible factor (HIF)-3alpha expression in lung epithelial cells: characterization and comparison with HIF-1alpha. Cell Res. 2006 Jun;16(6):548–58.

105.	Smith TG, Robbins PA, Ratcliffe PJ. The human side of hypoxia-inducible factor. Br J Haematol. 2008 May;141(3):325–34.

106.	Richardson SM, Knowles R, Tyler J, Mobasheri A, Hoyland JA. Expression of glucose transporters GLUT-1, GLUT-3, GLUT-9 and HIF-1alpha in normal and degenerate human intervertebral disc. Histochem Cell Biol. 2008 Apr;129(4):503–11.

107.	Vannucci SJ, Maher F, Simpson IA. Glucose transporter proteins in brain: delivery of glucose to neurons and glia. Glia. 1997 Sep;21(1):2–21.

108.	Takeda J, Kayano T, Fukomoto H, Bell GI. Organization of the human GLUT2 (pancreatic beta-cell and hepatocyte) glucose transporter gene. Diabetes. 1993 May;42(5):773–7.

109.	Heather LC, Pates KM, Atherton HJ, Cole MA, Ball DR, Evans RD, et al. Differential translocation of the fatty acid transporter, FAT/CD36, and the glucose transporter, GLUT4, coordinates changes in cardiac substrate metabolism during ischemia and reperfusion. Circ Heart Fail. 2013 Sep 1;6(5):1058–66.

110.	Kim J, Tchernyshyov I, Semenza GL, Dang CV. HIF-1-mediated expression of pyruvate dehydrogenase kinase: a metabolic switch required for cellular adaptation to hypoxia. Cell Metab. 2006 Mar;3(3):177–85.

111.	Kim BY, Kim H, Cho EJ, Youn HD. Nur77 upregulates HIF-alpha by inhibiting pVHL-mediated degradation. Exp Mol Med. 2008 Feb 29;40(1):71–83.

112.   Choi J-W, Park SC, Kang GH, Liu JO, Youn H-D. Nur77 activated by hypoxia-inducible factor-1alpha overproduces proopiomelanocortin in von Hippel-Lindau-mutated renal cell carcinoma. Cancer Res. 2004 Jan 1;64(1):35–9.

113.   Roe J-S, Youn H-D. The positive regulation of p53 by the tumor suppressor VHL. Cell Cycle Georget Tex. 2006 Sep;5(18):2054–6.

114.   Fels DR, Koumenis C. HIF-1alpha and p53: the ODD couple? Trends Biochem Sci. 2005 Aug;30(8):426–9.

115.   Berndt JD, Moon RT, Major MB. Beta-catenin gets jaded and von Hippel-Lindau is to blame. Trends Biochem Sci. 2009 Mar;34(3):101–4.

116.   Hergovich A, Lisztwan J, Thoma CR, Wirbelauer C, Barry RE, Krek W. Priming-dependent phosphorylation and regulation of the tumor suppressor pVHL by glycogen synthase kinase 3. Mol Cell Biol. 2006 Aug;26(15):5784–96.

117.   Heiner M, Koch I, Will J. Model validation of biological pathways using Petri nets--demonstrated for apoptosis. Biosystems. 2004 Jul;75(1-3):15–28.

118.   Bortfeldt RH, Schuster S, Koch I. Exhaustive analysis of the modular structure of the spliceosomal assembly network: a petri net approach. Stud Health Technol Inform. 2011;162:244–78.

119.   Heiner M. Understanding Network Behavior by Structured Representations of Transition Invariants. In: Condon A, Harel D, Kok JN, Salomaa A, Winfree E, editors. Algorithmic Bioprocesses [Internet]. Springer Berlin Heidelberg; 2009 [cited 2015 Jan 7]. p. 367–89. Available from: http://link.springer.com/chapter/10.1007/978-3-540-88869-7_19

120.   Zhu JL, Kaytor EN, Pao CI, Meng XP, Phillips LS. Involvement of Sp1 in the transcriptional regulation of the rat insulin-like growth factor-1 gene. Mol Cell Endocrinol. 2000 Jun;164(1-2):205–18.

121.   Okabe T, Nawata H. [Functional role of nur77 family in T-cell apoptosis and stress response]. Nihon Rinsho Jpn J Clin Med. 1998 Jul;56(7):1734–8.

122.   Murphy EP, Conneely OM. Neuroendocrine regulation of the hypothalamic pituitary adrenal axis by the nurr1/nur77 subfamily of nuclear receptors. Mol

196

Endocrinol Baltim Md. 1997 Jan;11(1):39–47.

123. Tabernero J. The role of VEGF and EGFR inhibition: implications for combining anti-VEGF and anti-EGFR agents. Mol Cancer Res MCR. 2007 Mar;5(3):203–20.

124. Masiero A, Aufiero S, Minervini G, Moro S, Costa R, Tosatto SCE. Evaluation of the steric impact of flavin adenine dinucleotide in Drosophila melanogaster cryptochrome function. Biochem Biophys Res Commun. 2014 Aug 8;450(4):1606–11.

125. Dym O, Eisenberg D. Sequence-structure analysis of FAD-containing proteins. Protein Sci Publ Protein Soc. 2001 Sep;10(9):1712–28.

126. Humphrey W, Dalke A, Schulten K. VMD: visual molecular dynamics. J Mol Graph. 1996 Feb;14(1):33–8, 27–8.

127. Van den Berg PAW, Feenstra KA, Mark AE, Berendsen HJC, Visser AJWG. Dynamic Conformations of Flavin Adenine Dinucleotide: Simulated Molecular Dynamics of the Flavin Cofactor Related to the Time-Resolved Fluorescence Characteristics. J Phys Chem B. 2002 Aug;106(34):8858–69.

128. Schreuder HA, Mattevi A, Obmolova G, Kalk KH, Hol WGJ, van der Bolt FJT, et al. Crystal Structures of Wild-Type p-Hydroxybenzoate Hydroxylase Complexed with 4-Aminobenzoate, 2,4-Dihydroxybenzoate, and 2-Hydroxy-4-aminobenzoate and of the Tyr222Ala Mutant Complexed with 2-Hydroxy-4-aminobenzoate. Evidence for a Proton Channel and a New Binding Mode of the Flavin Ring. Biochemistry (Mosc). 1994 Aug;33(33):10161–70.

129. Polticelli F, Basran J, Faso C, Cona A, Minervini G, Angelini R, et al. Lys300 plays a major role in the catalytic mechanism of maize polyamine oxidase. Biochemistry (Mosc). 2005 Dec 13;44(49):16108–20.

130. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem. 2004 Oct;25(13):1605–12.

131. Glykos NM. Software news and updates. Carma: a molecular dynamics

analysis program. J Comput Chem. 2006 Nov 15;27(14):1765–8.

132.   Tsoulos IG, Stavrakoudis A. Eucb: A C++ program for molecular dynamics trajectory analysis. Comput Phys Commun. 2011 Mar;182(3):834–41.

133.   Walsh I, Minervini G, Corazza A, Esposito G, Tosatto SCE, Fogolari F. Bluues server: electrostatic properties of wild-type and mutated protein structures. Bioinforma Oxf Engl. 2012 Aug 15;28(16):2189–90.

134.   Notredame C, Higgins DG, Heringa J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol. 2000 Sep 8;302(1):205–17.

135.   Czarna A, Berndt A, Singh HR, Grudziecki A, Ladurner AG, Timinszky G, et al. Structures of Drosophila cryptochrome and mouse cryptochrome1 provide insight into circadian function. Cell. 2013 Jun 6;153(6):1394–405.

136.   Minervini G, Masiero A, Potenza E, Tosatto SCE. Structural protein reorganization and fold emergence investigated through amino acid sequence permutations. Amino Acids. 2015 Jan;47(1):147–52.

137.   Li Y, Zhao Y, Hatfield S, Wan R, Zhu Q, Li X, et al. Dipeptide seryl-histidine and related oligopeptides cleave DNA, protein, and a carboxyl ester. Bioorg Med Chem. 2000 Dec;8(12):2675–80.

138.   Plankensteiner K, Righi A, Rode BM. Glycine and Diglycine as Possible Catalytic Factors in the Prebiotic Evolution of Peptides. Orig Life Evol Biosph. 2002 Jun 1;32(3):225–36.

139.   Zhang B, Cech TR. Peptide bond formation by in vitro selected ribozymes. Nature. 1997 Nov 6;390(6655):96–100.

140.   Johnston WK, Unrau PJ, Lawrence MS, Glasner ME, Bartel DP. RNA-catalyzed RNA polymerization: accurate and general RNA-templated primer extension. Science. 2001 May 18;292(5520):1319–25.

141.   Berger G. Deterministic hypotheses on the origin of life and of its reproduction. Med Hypotheses. 2003 Dec;61(5-6):586–92.

142.   Weiss O, Jiménez-Montaño MA, Herzel H. Information content of protein sequences. J Theor Biol. 2000 Oct 7;206(3):379–86.

143.   Minervini G, Evangelista G, Villanova L, Slanzi D, De Lucrezia D, Poli I, et al. Massive non-natural proteins structure prediction using grid technologies. BMC Bioinformatics. 2009;10 Suppl 6:S22.

144.   Lavelle DT, Pearson WR. Globally, unrelated protein sequences appear random. Bioinforma Oxf Engl. 2010 Feb 1;26(3):310–8.

145.   Pearl F, Todd A, Sillitoe I, Dibley M, Redfern O, Lewis T, et al. The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. Nucleic Acids Res. 2005 Jan 1;33(Database issue):D247–51.

146.   Turian G. Origin of life. II. From prebiotic replicators to protocells. Arch Sci C r Séances Société Édités Par Société Phys Hist Nat Genève. 1999 Aug;52(2):101–9.

147.   Rohl CA, Strauss CEM, Misura KMS, Baker D. Protein structure prediction using Rosetta. Methods Enzymol. 2004;383:66–93.

148.   Walsh I, Martin AJM, Di Domenico T, Tosatto SCE. ESpritz: accurate and fast prediction of protein disorder. Bioinforma Oxf Engl. 2012 Feb 15;28(4):503–9.

149.   Walsh I, Seno F, Tosatto SCE, Trovato A. PASTA 2.0: an improved server for protein aggregation prediction. Nucleic Acids Res. 2014 Jul;42(Web Server issue):W301–7.

150.   Bradley P, Malmström L, Qian B, Schonbrun J, Chivian D, Kim DE, et al. Free modeling with Rosetta in CASP6. Proteins. 2005;61 Suppl 7:128–34.

151.   Das R, Baker D. Macromolecular modeling with rosetta. Annu Rev Biochem. 2008;77:363–82.

152.   Martin-Visscher LA, Gong X, Duszyk M, Vederas JC. The three-dimensional structure of carnocyclin A reveals that many circular bacteriocins share a common structural motif. J Biol Chem. 2009 Oct 16;284(42):28674–81.

153.   Chiarabelli C, Vrijbloed JW, De Lucrezia D, Thomas RM, Stano P, Polticelli F, et al. Investigation of de novo totally random biosequences, Part II: On the folding frequency in a totally random library of de novo proteins obtained by phage display. Chem Biodivers. 2006 Aug;3(8):840–59.

154. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997 Sep 1;25(17):3389–402.

155. Iakoucheva LM, Brown CJ, Lawson JD, Obradović Z, Dunker AK. Intrinsic disorder in cell-signaling and cancer-associated proteins. J Mol Biol. 2002 Oct 25;323(3):573–84.

156. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol. 1995 Apr 7;247(4):536–40.

157. Steel M, Penny D. Origins of life: Common ancestry put to the test. Nature. 2010 May 13;465(7295):168–9.

158. Theobald DL. A formal test of the theory of universal common ancestry. Nature. 2010 May 13;465(7295):219–22.

159. Commeyras A, Collet H, Boiteau L, Taillades J, Vandenabeele-Trambouze O, Cottet H, et al. Prebiotic synthesis of sequential peptides on the Hadean beach by a molecular engine working with nitrogen oxides as energy sources. Polym Int. 2002;51(7):661–5.

160. Kauffman S. Molecular autonomous agents. Philos Trans R Soc Lond Math Phys Eng Sci. 2003 Jun 15;361(1807):1089–99.

161. Labean TH, Butt TR, Kauffman SA, Schultes EA. Protein folding absent selection. Genes. 2011;2(3):608–26.

162. Scaini MC, Minervini G, Elefanti L, Ghiorzo P, Pastorino L, Tognazzo S, et al. CDKN2A unclassified variants in familial malignant melanoma: combining functional and computational approaches for their assessment. Hum Mutat. 2014 Jul;35(7):828–40.

163. Das P, Kang S-G, Temple S, Belfort G. Interaction of Amyloid Inhibitor Proteins with Amyloid Beta Peptides: Insight from Molecular Dynamics Simulations. PloS One. 2014;9(11):e113041.

164. Liu MS, Wang D, Morimoto H, Yim H, Irving AT, Williams BRG, et al.

Molecular dynamics reveal a novel kinase-substrate interface that regulates protein translation. J Mol Cell Biol. 2014 Nov 17;

165. Betz RM, Walker RC. Paramfit: Automated optimization of force field parameters for molecular dynamics simulations. J Comput Chem. 2014 Nov 21;

166. Karp JM, Erylimaz E, Cowburn D. Correlation of chemical shifts predicted by molecular dynamics simulations for partially disordered proteins. J Biomol NMR. 2014 Nov 22;

167. Varkuti BH, Yang Z, Malnasi-Csizmadia A. Structural model of weak-binding actomyosin in the prepowerstroke state. J Biol Chem. 2014 Nov 21;

168. Doncheva NT, Assenov Y, Domingues FS, Albrecht M. Topological analysis and interactive visualization of biological networks and protein structures. Nat Protoc. 2012 Apr;7(4):670–85.

169. Pasi M, Tiberti M, Arrigoni A, Papaleo E. xPyder: A PyMOL Plugin To Analyze Coupled Residues and Their Networks in Protein Structures. J Chem Inf Model. 2012 Jul 23;52(7):1865–74.

170. Tiberti M, Invernizzi G, Lambrughi M, Inbar Y, Schreiber G, Papaleo E. PyInteraph: a framework for the analysis of interaction networks in structural ensembles of proteins. J Chem Inf Model. 2014 May 27;54(5):1537–51.

171. Vijay-Kumar S, Bugg CE, Cook WJ. Structure of ubiquitin refined at 1.8 A resolution. J Mol Biol. 1987 Apr 5;194(3):531–44.

172. Wang Y, Amegbey G, Wishart DS. Solution structures of reduced and oxidized bacteriophage T4 glutaredoxin. J Biomol NMR. 2004 May;29(1):85–90.

173. Weaver LH, Matthews BW. Structure of bacteriophage T4 lysozyme refined at 1.7 A resolution. J Mol Biol. 1987 Jan 5;193(1):189–99.

174. Jain AK. Data clustering: 50 years beyond K-means. Pattern Recognit Lett. 2010 Jun 1;31(8):651–66.

175. Fariselli P, Olmea O, Valencia A, Casadio R. Prediction of contact maps with neural networks and correlated mutations. Protein Eng. 2001 Nov 1;14(11):835–43.

176. Ralat LA, Kalas V, Zheng Z, Goldman RD, Sosnick TR, Tang W-J. Ubiquitin is a novel substrate for human insulin-degrading enzyme. J Mol Biol. 2011 Feb 25;406(3):454–66.

177. Plechanovová A, Jaffray EG, Tatham MH, Naismith JH, Hay RT. Structure of a RING E3 ligase and ubiquitin-loaded E2 primed for catalysis. Nature. 2012 Sep 6;489(7414):115–20.

178. Huang DT, Walden H, Duda D, Schulman BA. Ubiquitin-like protein activation. Oncogene. 2004 Mar 15;23(11):1958–71.

179. Malito E, Hulse RE, Tang W-J. Amyloid beta-degrading cryptidases: insulin degrading enzyme, presequence peptidase, and neprilysin. Cell Mol Life Sci CMLS. 2008 Aug;65(16):2574–85.

180. Varadi M, Kosol S, Lebrun P, Valentini E, Blackledge M, Dunker AK, et al. pE-DB: a database of structural ensembles of intrinsically disordered and of unfolded proteins. Nucleic Acids Res. 2014 Jan;42(Database issue):D326–35.

181. Monzon AM, Juritz E, Fornasari MS, Parisi G. CoDNaS: a database of conformational diversity in the native state of proteins. Bioinforma Oxf Engl. 2013 Oct 1;29(19):2512–4.

182. Moussa SH, Kuznetsov V, Tran TAT, Sacchettini JC, Young R. Protein determinants of phage T4 lysis inhibition. Protein Sci Publ Protein Soc. 2012 Apr;21(4):571–82.

183. Nikkola M, Gleason FK, Eklund H. Reduction of mutant phage T4 glutaredoxins by Escherichia coli thioredoxin reductase. J Biol Chem. 1993 Feb 25;268(6):3845–9.

184. Zhang X, Wozniak JA, Matthews BW. Protein Flexibility and Adaptability Seen in 25 Crystal Forms of T4 Lysozyme. J Mol Biol. 1995 Jul 21;250(4):527–52.

185. Grant BJ, Rodrigues APC, ElSawy KM, McCammon JA, Caves LSD. Bio3d: an R package for the comparative analysis of protein structures. Bioinforma Oxf Engl. 2006 Nov 1;22(21):2695–6.

186. Komander D. The emerging complexity of protein ubiquitination. Biochem

Soc Trans. 2009 Oct;37(Pt 5):937–53.