Universitat
de Barcelona

Protein dynamics studied by coarse-grained and atomistic theoretical approaches

Laura Orellana

U
UNIVERSITAT DE BARCELONA
B

PROTEIN DYNAMICS STUDIED BY COARSE-GRAINED AND ATOMISTIC THEORETICAL APPROACHES

Laura Orellana

B.S. in Biochemistry with Honors, University of Barcelona, 2005

M.S. in Protein Structure and Function, Autonomous University of Barcelona, 2007

M.S. in Biophysics, University of Barcelona, 2010

Submitted to the Faculty of
Physics in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

University of Barcelona

2014

This page intentionally left blank

PHD PROGRAM IN PHYSICS

DEPARTAMENT OF FUNDAMENTAL PHYSICS

UNIVERSITY OF BARCELONA (UB)

&

INSTITUTE FOR RESEARCH IN BIOMEDICINE
(IRB BARCELONA)

PhD advisor: Prof. Modesto Orozco

Tutor: Dr. Matteo Palassini

This page intentionally left blank

Let us fight for a World of Reason!!

A world where Science and Progress
Will lead to all men's Happiness

Charlie Chaplin

***The Great Dictator* (1940)**

The Cosmos is all that is or was or ever will be.

Our feeblest contemplations of the Cosmos stir us -- there is a tingling in the spine, a catch in the voice, a faint sensation, as if a distant memory, of falling from a height.

We know we are approaching the greatest of mysteries.

Carl Sagan

***Cosmos* (1980)**

This page intentionally left blank

Al meu company, Johan, i a la meva mare.

Als que van marxar i als que venen.

This page intentionally left blank

PROTEIN DYNAMICS AND FUNCTION STUDIED BY COARSE-GRAINED AND ATOMISTIC THEORETICAL APPROACHES

Laura Orellana, Ph.D.
University of Barcelona, 2014

Protein structure, dynamics and function, are inseparable in order to understand the mechanisms of Life at the molecular level. From the structure comes dynamics, and from dynamics many, if not all, protein functions. Usually functional motions operate at timescales and conditions that are far beyond the limits of current experimental techniques. In order to address rationally the complex interplay between these three aspects of proteins, we will deepen into and compare both coarse-grained and atomistic simulation methods. In the first part of this thesis, coarse-grained Elastic Network Models are compared with Molecular Dynamics and experimental flexibility in order to obtain a more accurate representation of protein dynamics. We propose a novel ENM algorithm based on local chain topology, which renders results close to atomistic simulation methods. In the second part of the thesis, the novel method is used to detect dynamical hot spots in the Tyrosine-Kinase HER1, revealing why a number of oncogenic mutations are accumulated at hinge interdomain regions. Near-microsecond long simulations of the detected dynamical mutants are presented which unravel a critical intermediate in the large-scale conformational change that activates the protein. Finally, we explore the validity of the ENM potentials in a number of applications, from sampling transition pathways to understanding how correlated motions in secondary structures transmit information or why dynamics is tightly related with the connectivity of residue networks. The present thesis demonstrates that functional motions are encoded in shape and local-topology, and that perturbations hitting key regions can shift protein dynamics, suggesting a novel oncogenic and evolutive mechanism.

This page intentionally left blank

DINÀMICA I FUNCIÓ DE PROTEÏNES: UN ESTUDI AMB MÈTODES TEÒRICS ATOMÍSTICS I DE BAIXA RESOLUCIÓ

Laura Orellana, Ph.D.
University of Barcelona, 2014

L'estructura, la dinàmica i la funció de les proteïnes són inseparables alhora d'entendre els mecanismes de la vida al nivell molecular. La estructura determina la dinàmica, i és la dinàmica la que decideix la majoria, si no totes, les funcions de les proteïnes. Molt sovint els moviments funcionals operen en escales de temps i condicions que desafien els límits de les tècniques experimentals actuals. Per tal d'entendre racionalment la complexa interrelació existent entre aquests tres aspectes de les proteïnes, aprofundirem i compararem diferents mètodes de simulació, atomístics i també de baixa resolució. En la primera part d'aquesta tesis, els models de baixa resolució anomenats de "Xarxa Elàstica" es comparen amb Dinàmica Molecular i flexibilitat experimental per obtenir una representació més acurada de la dinàmica de les proteïnes. Proposem un nou algoritme de xarxa elàstica basat en la topologia local de la cadena, que proporciona resultats molt pròxims als mètodes de simulació atomístics. En la segona part de la tesis, el nou mètode s'utilitza per detectar punts calents per la dinàmica de la Tirosina Cinasa HER1, revelant perquè nombroses mutacions oncogèniques s'acumulen en regions interdomini clau. Presentem simulacions dels mutants dinàmics a prop de l'escala de microsegons que revelen un intermediari crític en el canvi conformacional de gran escala que activa la proteïna. Finalment, analitzem la validesa dels potencials de xarxa elàstica en una sèrie d'aplicacions: des d'explorar possibles camins per realitzar transicions conformacionals, a entendre com els moviments correlacionats de les estructures secundàries transmeten informació, o perquè la dinàmica esta estretament lligada a la connectivitat de la xarxa d'aminoàcids. La present tesis demostra que els moviments funcionals es troben codificats en la forma i la topologia local, i que les pertorbacions en regions crítiques poden alterar la dinàmica de les proteïnes, suggerint un nou mecanisme oncogènic i evolutiu.

This page intentionally left blank

Table of Contents

This page intentionally left blank

List of publications derived from this thesis

1. Orellana L., and Orozco M. (2014). **Elastic Network-Brownian Dynamics transition pathways in large-scale conformational changes**. *(in preparation)*

   CONTRIBUTIONS: L.O. proposed research, designed and performed all calculations and wrote the paper with M.O.

2. Orellana L., López-Blanco J.R., Chacón P., Orozco M. (2014) **Exploring the link between residue network properties and protein dynamical behavior by fast dihedral NMA**. *(in preparation)*

   CONTRIBUTIONS: L.O. proposed research, designed and performed all calculations and wrote the paper with M.O.

JCTC Journal of Chemical Theory and Computation
SEPTEMBER 2010 Volume 6 Number 9 pubs.acs.org/JCTC
ACS Publications
www.acs.org

3. Orellana L., Hospital A. and Orozco M. (2013). **Unraveling the dynamics of epidermal growth factor receptor through oncogenic mutations.** *JACS (submitted)*

   CONTRIBUTIONS: L.O. proposed research, designed and performed NMA calculations and MD analysis and wrote the paper with M.O.

4. Fenwick R.B.*, Orellana L.*, Esteban-Martin S., Salvatella X. and Orozco M. (2013) **Correlated Motions in β-sheets are an Inherent Property of Proteins.** *(Nature Communications, 2nd review)* **Equal contribution*

   CONTRIBUTIONS: L.O. contributed designing research, performing all MD and NMA calculations and writing the paper.

5. Sfriso P., Emperador A., Orellana L., Hospital A., Gelpí J.L., Orozco M. **Finding Conformational Transition Pathways from Discrete Molecular Dynamics Simulations.** *J. Chem. Theory Comput.* (2012) 8 (11): p 4707–4718

   CONTRIBUTIONS: L.O. contributed in NMA calculations to guide dMD

6. Orellana L., Rueda M., Ferrer-Costa C., López-Blanco J.R., Chacón P. and Orozco M. **Approaching Elastic Network Models to Molecular Dynamics Flexibility.** *J. Chem. Theory Comput.* (2010) 6: p2910–2923. Featured Cover Paper: JCTC September 2010 Cover

   CONTRIBUTIONS: L.O. performed all calculations and wrote the paper with M.O.

7. Orozco M., Orellana L., Hospital A., Naganathan A.N., Emperador A., Carrillo O. and Gelpí J.L.. (2010) **Coarse Grained Representation of Protein Flexibility. Foundations, successes and shortcomings. In "Computational Chemistry Methods in Structural Biology".** *Advances in Protein Chemistry and Structural Biology*, vol 80. Ed. C.Christov. Elsevier.

   CONTRIBUTIONS: L.O. contributed writing and revising the paper.

8. Camps J., Carrillo O., Emperador A., Orellana L., Hospital A., Rueda M., Gelpi JL. & Orozco M. **"FlexServ: An integrative tool for the Analysis of protein Flexibility"** *Bioinformatics* (2009) 25 (1): p1709–1710. *CONTRIBUTIONS: L.O. contributed with NMA implementation*

This page intentionally left blank

List of Tables

Chapter 1

Chapter 3

Chapter 4

Chapter 5

List of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

List of Abbreviations

Abbreviation	**Full Name**
ANM	*Anisotropic Network Model*
BD	*Brownian Dynamics*
CG	*Coarse-graining*
DMD	*Discrete Molecular Dynamics*
DIMS	*Dynamic Importance Sampling*
ED	*Essential Dynamics*
ED-ENM	*Essential Dynamics-derived Elastic Network Model*
EGFR	*Epidermal Growth Factor Receptor*
ENM	*Elastic Network Model*
GNM	*Gaussian Network Model*
HER	*Human EGFR*
MD	*Molecular Dynamics*
NMA	*Normal Mode Analysis*
NMR	*Nuclear Magnetic Resonance*
PCA	*Principal Component Analysis*

List of Parameters

Parameter	**Full Name**	**Value (units)**
$\mathbf{k_BT}$	Thermal factor	0.59 kcal/mol

“All things are in motion”

Attributed to Heraclitus

1- Introduction

This page intentionally left blank

Chapter 1 Introduction: Protein Structure, Function and Motion

In this first chapter, we introduce the problem of protein flexibility and how it is related to biological shape and function. Then we outline briefly the two main theoretical approaches to explore protein motions: atomistic and coarse-grained, which will be discussed throughout this thesis, from the methodological basis **(Chapter 2)** to their comparison with experimental data **(Chapters 3)**, and how they can be applied to dissect the function of a protein in disease **(Chapter 4)**, or to solve a diverse array of problems in structural biology and dynamics **(Chapter 5).**

1.1 Protein Flexibility and Function

Dynamics is the key to understand function

In the late XIXth century, fast photography techniques were developed allowing for the first time to capture sequences of motion, such as the one entitled *"Sallie Gardner at a Gallop"* (*Figure 1. 1*, *left*), one of the first movies in history. The author of this amazing photographic series was **Edwaeard Muybridge (1840-1904)**, a visionary photographer interested in body motion, which helped to develop fast photography. Using his improved camera, he took the *"Horse in motion"* series and won a bet to figure out whether all four feet of a horse were off the ground at the same time while galloping. Looking at these beautiful sequences of images, it is evident that not only the shape of the horse's body, but also the coordinated motion of the legs, has been designed to run. Evolution selects the shapes best suited to perform certain functional motions – from running to swimming or flying – and the same principle is valid for horses, biomolecules, and all living matter (Bejan & Lorente, 2010). Protein structural changes are encoded in the sequence, which determines the overall fold and in the end, the intrinsic motions and the resulting biological function. These functional movements exist by evolutionary selection which guide protein sequences evolution, not only to adopt certain structures, but to favor dynamical properties required for function (Velázquez-Muriel et al., 2009). Each structure can only sample a limited part of the conformational space, and thus, displays a reduced set of elemental motions; the shape determines the conformations/motions, and the motions determine the function. Therefore, when two sequences have even a remote but sizeable sequence similarity (around 30 %), it is very probable that they will share a common fold, and even more, a similar flexibility pattern and function.

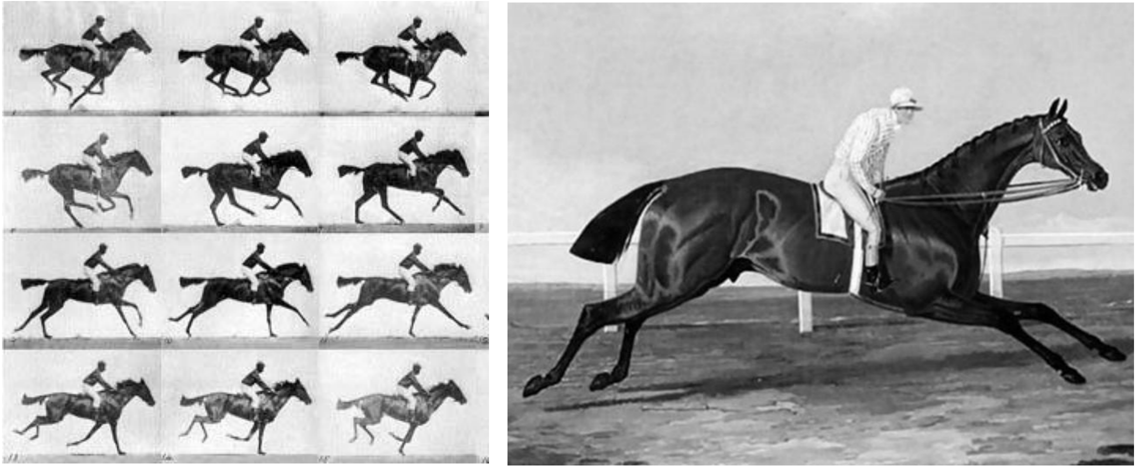


Figure 1. 1 Edwaeard Muybridge "Sallie Gardner at gallop" and a XIXth Century equestrian painting. Thanks to the development of a fast camera, the British photographer Edwaeard Muybridge won a bet to figure out whether all four feet of a horse are off the ground simultaneously while galloping. Until Muybridge took multiple snapshots of running horses ("*The Horse in Motion"* series, *left*) nobody understood correctly the gallop mechanism. The detailed mechanics of this motion is too fast for the human eye, and thus it's not surprising that all artistic representations of running horses before the 20th century were remarkably wrong, with a "flying gallop" pose (note that at the moment that no hoof touches the ground, the horse's legs are gathered together, not splayed, *right*).

In spite of the relevance of dynamics for evolution, we usually think of proteins as fixed, rigid structures, like the flat, static images that appear in every biochemistry book - the so-called *native* conformations nicely trapped in crystals. However, proteins are highly dynamical molecules, ever-changing and continuously exploring the conformational space - in some sense, they behave as *living* entities (Henzler-Wildman & Kern, 2007). If biomolecules had rigid structures, they could not perform any useful task in the cell other than mechanical support as mere building blocks - such as collagen, or keratin. Proteins need to change their form, to have a certain degree of conformational flexibility to act in their environment. This is obvious for motor proteins such as myosin, responsible of muscle contraction (see *Figure 1. 2*), or kinesin, involved in the busy vesicle trafficking inside the cells. But even processes like enzyme catalysis or protein-protein binding involve important structural changes (Karplus & Kuriyan, 2005; Karplus & McCammon, 2002). Almost any interaction between a protein and another molecule is associated with smaller or larger conformational changes: ligand recognition, signal transduction, assembly of multiprotein machines or allosteric regulation. Extreme cases of flexibility are the **molten globule** (Naganathan & Orozco, 2011) and **intrinsically disordered proteins** (Babu, Van Der Lee, De Groot, & Gsponer, 2011), in which the traditional concept of shape is lost and function arises from conformational freedom.

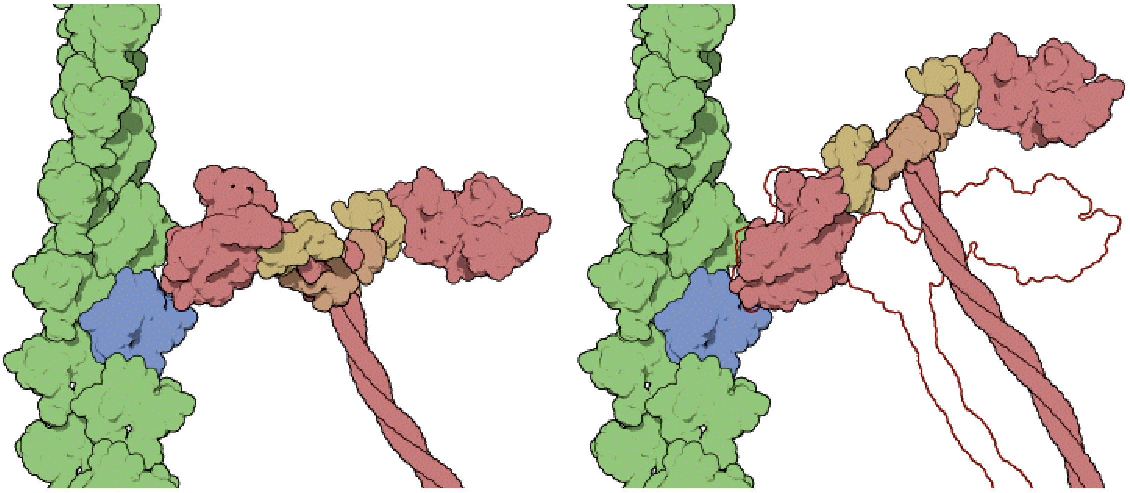


Figure 1. 2 Protein motion is necessary for biological function.
ATP hydrolysis provides the energy to trigger conformational changes in the molecular motor myosin, responsible for motion in all muscular cells. When ATP is cleaved, myosin adopts a bent, flexed form, like in the structure on the left (shown in pink and ochre, PDB *1br1*). This prepares myosin for the power stroke. The flexed myosin then grabs the actin filament (shown in green and blue, from PDB *1atn*), and it is the release of phosphate that snaps the motor into a straight "rigor" form, as shown on the right (PDB *2mys*). This *power stroke* pushes the myosin molecule along the actin filament. When finished, the remaining ADP is replaced by a new ATP, the myosin lets go of the actin filament and it's ready for the next stroke.

Under the thermal noise and multiple influences from the cell environment, each one of the atoms of a protein is departing at every femtosecond from its average position, only subject to the constraints imposed by chemical interactions, which dictate the overall protein shape. Some constraints are stronger that others (i.e. covalent bonds) but as a whole they allow the structure to display a range of typical motions near equilibrium, to visit a finite but wide range of thermally-accessible states. However, the most powerful structural technique available, X-ray diffraction, provides only a single frozen picture from a conformational ensemble in terms of average positions. These coordinates provide as much information as an isolated view from a resting horse, and no dynamic mechanisms can be inferred from it. Not surprisingly, all pictorial representations of galloping horses before the advent of fast cameras were radically wrong (see *Figure 1. 1, right*). Although cavemen were better observers than all modern people, the fine details of these complex series of motions were also too fast for their gaze (Horvath, Farkas, Boncz, Blaho, & Kriska, 2012). Up to now there is no experimental technique that can provide us with a video of a moving, *real*, protein structure, and in the best of the cases, we have a couple of photos from different poses (*conformations*). As Muybridge's history shows, it is extremely difficult to reconstruct the mechanics of motion from static views. However, we can always try to *compute* the motions using the laws of physics; ultimately, the key to truly understand the link between molecular shapes, motion and function is to explain their action in terms of physical forces. In other words, simulations provide the most rational approach to dissect protein dynamics.

The evolving view on protein function: from aperiodic crystals to disordered structures

The concept of proteins as static entities appeared long before any structural data was available; actually, it can be traced back to the *''key-lock''* hypothesis proposed at the end of the XIXth century by **Emil Fischer** (Fischer, 1894) to explain enzymatic activity in terms of what we would call today a *"rigid docking"* – a well-defined active site complementary to a substrate's fixed shape. This idea persisted throughout the first half of the past century, even in revolutionary physicists such as **Erwin Schrödinger** (Schrödinger, 1992) which conceived the molecules of life as some sort of *"aperiodic crystals"*. However, this view of proteins as static and rigid objects proved to be insufficient to explain an increasing body of experimental data, especially in the field of enzymology. The first to envision a structural explanation for the function of a protein, Hemoglobin, in terms of ligand-mediated conformational changes, was the brilliant **Linus Pauling**, almost three decades before the first X-ray structures were solved (Pauling, 1935). The fundamental link between the protein structure, and its dynamics and biological function, became evident when the first atomic-resolution models were obtained by **John Kendrew** and **Max Perutz** (Kendrew et al., 1958). In the same year, **Daniel Koshland** (Koshland, 1958) introduced the *''induced fit''* theory and with it the idea of "structural changes" in proteins upon binding of a substrate; and a year later, **Linderstrom-Lang** and **Schellmann** (Linderstrom-Lang & Schellmann, 1959) proposed a breath-like continuous movement of the proteins, a fruitful idea experimentally probed during the following decades. When the crystallographic apo and holo structures of Lysozyme were solved (alone and in complex with inhibitors (Blake et al., 1965; Johnson & Phillips, 1965)) – the role of conformational flexibility for catalytic mechanisms, and for protein function in general, began to be widely accepted. The alternative view to the induced-fit theory of allostery, the *"conformational selection"* or **Monod-Wyman-Changeux model**, based on concerted transitions between preexisting protein conformations was proposed that same year (Monod, Wyman, & Changeux, 1965). The development of Molecular Dynamics (MD) simulations in the 70s (J A McCammon, Gelin, & Karplus, 1977) and Nuclear Magnetic Resonance (NMR) (Wuethrich, 1989) in the 80s confirmed the view that proteins are highly dynamic and that their different scales of motion - from femtosecond atom vibrations to second-long folding and unfolding processes - are fundamental for their functions. This dynamic picture has been confirmed by time-resolved crystallography (Hajdu et al., 2000), Förster resonance energy transfer (FRET) (Brunger, Strop, Vrljic, Chu, & Weninger, 2011; Heyduk, 2002), or neutron scattering (Bernadó, Mylonas, Petoukhov, Blackledge, & Svergun, 2007; Gabel et al., 2002; Putnam, Hammel, Hura, & Tainer, 2007), that provide precious information from functional conformational changes.

The problem of time and length scales in living matter: the gap between theory and experimentation

Despite the huge advances in the past few years, one of the main unresolved problems in biology is the huge gap existing between the time and length scales that can be addressed at the computational and experimental levels. The study of protein flexibility is still very challenging for experimental techniques and much information is derived from theoretical methods, **Molecular Dynamics (MD)** being the most rigorous one. However, whereas experimental approaches reach a maximal resolution of ms-µs in time, and a length order of nm, theoretical-computational methods allow simulating the microscopic level, but in general only on the order of 10^2 ns/10^1 nm, for small-to-middle sized proteins. Simulations in the microsecond timescale for large systems are still prohibitive in terms of computational cost, not to say the study of protein-protein or protein-membrane interactions, only available with specialized supercomputers (Arkhipov et al., 2013). Precisely, the most relevant interactions *in vivo*, such as concerted domain motions, docking or protein folding, occur in this blurred boundary between theory and experimentation: they are slow motions, and often involve great macromolecular complexes (see *Table 1. 1* below). In these fundamental processes, the number of degrees of freedom is still many orders of magnitude beyond the frontier we can reach in the present in length and time scales.

Table 1. 1 Time scale and Amplitude of Biomolecular motions.

TIME SCALE	AMPLITUDE	DESCRIPTION
short femto, pico **10^{-15} - 10^{-12}s**	0.001 - 0.1 Å	**Local motions** (0.01 to 5 Å, 10^{-15} to 10^{-1} s) - Atomic fluctuations: bond stretching, angle bending, dihedral motion - Side chain motions - Loop motions
medium pico, nano **10^{-12} - 10^{-9}s**	0.1 - 10 Å	**Rigid Body Motions** (1 to 10Å, 10^{-9} to 1s) - Helix motions - Collective motions: Subunit and Domain Motions (hinge bending)
long nano, micro **10^{-9} - 10^{-6}s**	1 - 100 Å	**Large- Scale Motions** (> 5Å, 10^{-7} to 10^{4} s) - Folding in small peptides - Helix coil transitions - Dissociation/Association
very long micro, second **10^{-6} - 10^{-1}s**	10 - 100 Å	- Folding and Unfolding

1.2 Overview: Molecular Dynamics and Coarse-Grained Models

The limitations of MD: coarse-graining to reach the mesoscopic scale

Molecular dynamics (MD) simulation, based on physical potentials which proceed from the rigorous formalism of molecular physics (see *below*), is the most powerful approach to model protein motions. Since the first protein simulation in the late 70s (J A McCammon et al., 1977), the size of the system that can be addressed by MD has been increasing gradually thanks to the better computational architectures, more efficient sampling techniques, and ever improved parameterization of the force-fields. In its current implementation the technique allows exploring dynamics in the multi-microsecond simulation range (Klepeis, Lindorff-Larsen, Dror, & Shaw, 2009), reaching the millisecond scale when special computers are used. However, the computational cost of MD still limits the accessible timescale, particularly for large molecular assemblies. Analysis of the current version of the protein databank (Berman et al., 2000) illustrates that most proteins with solved experimental structures have between 500 and 7000 atoms, but in some cases systems reach more than 16000 atoms (see *Figure 1. 3*). As experimental resolution techniques advance and larger protein assemblies are incorporated, the size-histogram is expected to displace more and more to the right.

Size is a major limitation for the theoretical simulation of proteins, but even more dramatic than the size-problem is the time-problem. Proteins are flexible, they move continuously at physiological temperatures, and as discussed above, although atomic vibrations happen in the nanosecond time scale, most biologically-relevant protein movements happen in the millisecond to second range. Each level of protein motion translates into the next, creating wider and slower movements. For example, local atomic vibrations are transmitted via hydrogen bond networks that make up secondary structures, creating higher amplitude motions; as we will see in **Chapter 5**, the coupled movements of interacting atoms in beta-sheet motifs create collective motions such as bending and twisting, which also play a role in higher order, collective movements of the protein related to function.

If we want to follow with atomistic detail the relevant motions, we have to start in the scale of femtosecond vibrations and beyond, and therefore, protein energy (and associated forces) has to be computed at least 10^{12} times for a modest millisecond of trajectory. For a typical 50000 atoms system, just the calculation of inter-atomic distances would require in the order of 10^{21} floating point operations, not far from the Avogadro number. In our MoDEL (***Molecular Dynamics Extended Library***) database of atomistic simulations (Meyer et al., 2010; Rueda, Ferrer-Costa, et al., 2007a) of representative PDB proteins in water, typical protein systems are in the order of 10000 to 50000 atoms, but some of them have more than 150000 atoms, i.e we are dealing with systems with up to *half a million degrees of freedom*. We will find a clear example in **Chapter 4**, where the simulation of the HER extracellular domain has required protein-solvent systems in the order of 200000-400000 atoms. Even more, if we are interested in studying more realistic protein-protein interactions, large macromolecular assemblies such as the ribosome, cell membranes, etc or diffusion or aggregation processes, sizes of simulated systems easily reach the many millions degrees of freedom, making atomistic simulation unattainable even for specialized supercomputers (Shaw et al., 2009). Therefore, simulations including all atoms and explicit solvent are only feasible for ns-µs motions, and to approach the larger, mesoscopic scale, where biologically relevant processes occur, the physical description of the system must change from the atomistic to a lower-resolution **Coarse-Grained (CG) model**.

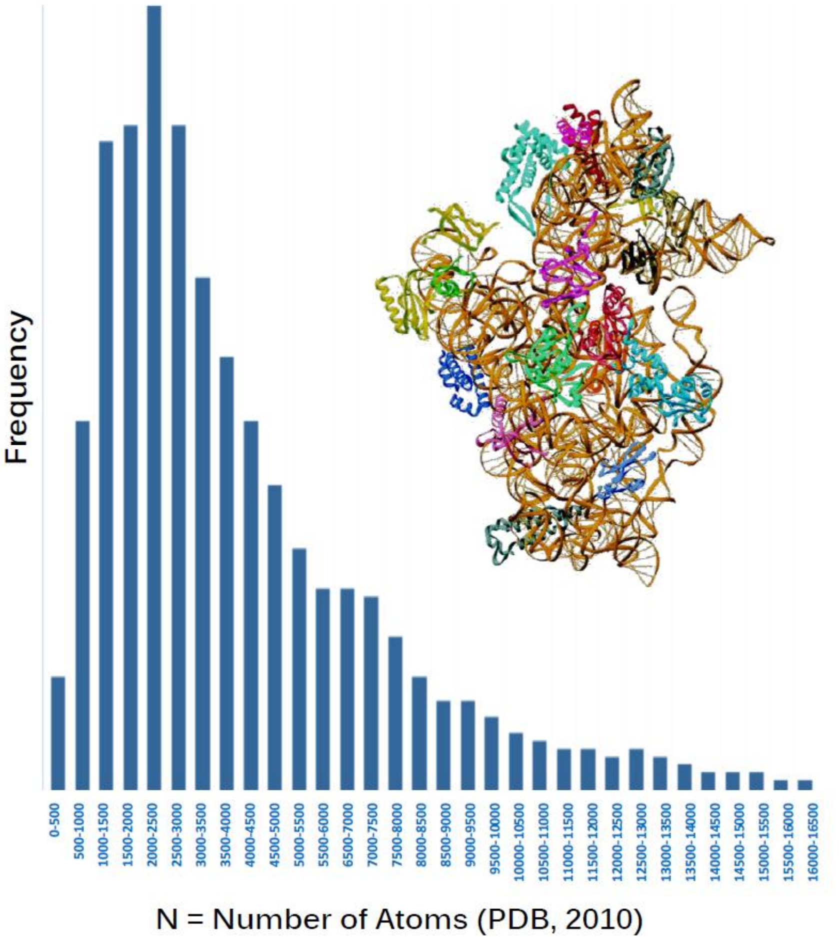


Figure 1. 3 Distribution of protein atoms in the Protein Data Bank (PDB).
Vertical histograms show the distribution of protein atoms in 2010 version. The prokaryotic ribosomes solved by the group of Ada Yonath (Nobel Prize in Chemistry, 2009) are among the largest high-resolution X-ray structures deposited in the PDB (right, 30S subunit of *T.termophilus*, 1FKA). Published in: Schluenzen et al; *Cell.* 2000, 102 (5): 615-23.

In physical modeling, coarse-graining consists in reducing the studied system to its main features, *i.e.* to simplify the description in order to be able to describe its behavior with a minimal number of parameters. Although renouncing to details, coarse-graining is not bad, as a master like the 17th century painter Diego Velázquez probably knew very well: he only needed a couple of fast strokes to capture the essential traits of a subject (see *Figure 1. 4*). Actually, this attitude is not only in the core of science but also constitutes a measure of the goodness of any scientific model; according to Occam's razor, "***Entities should not be multiplied unnecessarily***", and therefore, the simpler a model is, the better. If we want to include small-scale details, a high resolution view is needed – an atomistic portrait of the system, like a hyperrealistic Coello painting (see *Figure 1. 4*) – but if we deal with large-scale phenomena, simplification is not only useful, but necessary.

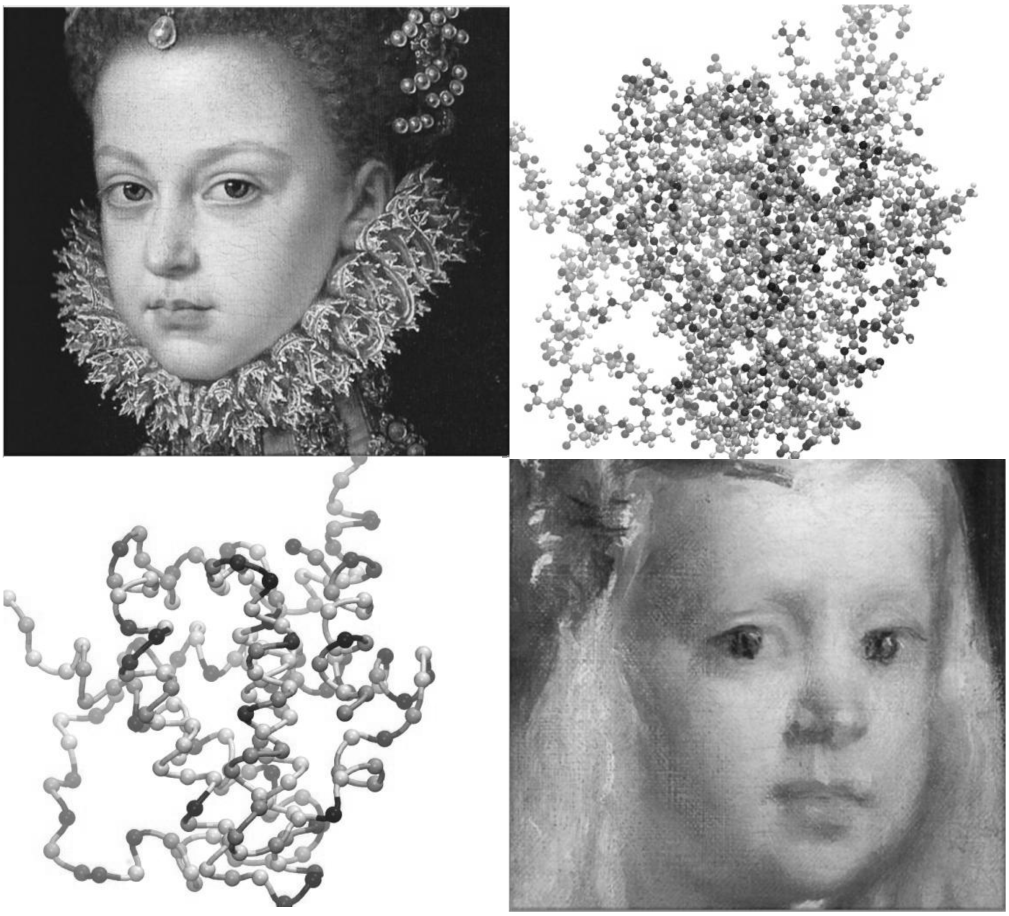


Figure 1. 4 Fine-grained and Coarse-grained views in Protein Structure and Art.
The atomistic description of proteins (*upper right corner*) can be compared to the hyperrealistic portraits of Spanish painter Claudio Coello (*upper left corner*), where the smallest details are registered – even the eyelashes of the girl. On the contrary, the impressionistic style of its contemporary Diego Velázquez (*lower right corner*) captures a face with few essential but blurred strokes; just as the coarse-graining of a protein reduces the complexity of a thousand of atoms system to the nude C-alpha carbon trace (*lower left corner*).

Fortunately, near-equilibrium protein dynamics is much simpler than one could expect; it can be described by a surprisingly low number of collective degrees of freedom. Since protein fundamental motions are often large scale-concerted displacements, there is no need to integrate the fine fluctuations of each atom to get the global picture of concerted motions of secondary structure elements and domains. When dealing with large biomolecular aggregates, experimentalists apply low resolution techniques such as Electron Microscopy or X-ray dispersion. Their computational counterpart is CG models, which simplify both the potentials and the structure description to extend the simulations near the experimental level. By eliminating fine details, computation is accelerated increasing many orders of magnitude the time and length scales available. However, one has to take into account that, in order to increase computer efficiency, there is a certain loss of accuracy and resolution of the data derived. Among CG methods, **Normal Mode Analysis (NMA)** (Case, 1994) stands out, being a classical technique based on simple harmonic potentials, which was developed to analyze the infrared spectra of simple atoms, but that has been proven to trace near equilibrium large-scale motions. Further coarse-graining of both the potential and the protein description, limited to the network of alpha-carbons, leads to the minimal **Elastic Network Models (ENMs)** (Atilgan et al., 2001; I Bahar, Atilgan, & Erman, 1997; Tirion, 1996), that will be analyzed in this thesis, both at the methodological **(Chapter 3)** and at the applications level **(Chapter 4)**, studying a biomedically relevant protein that undergoes a large functional transition. We will combine both atomistic simulations and coarse-grained methods to study protein dynamics, with a particular focus on large-scale conformational changes. As we will see, the fact that large-scale motions define the functional dynamics also facilitates the analysis of MD trajectories through **Essential Dynamics (ED)**, a technique that we will use repeatedly as reference. We will develop in detail the theory behind these methods in the next chapter **(Chapter 2)**.

“Everything that is living can be understood in terms of the jiggling and wiggling of atoms.”

Richard Feynman

2- Theoretical frameworks

This page intentionally left blank

Chapter 2 Theoretical Frameworks: atomistic versus coarse-grained

In this Chapter, we will explain the basis of molecular simulations of proteins, and discuss in detail the two main approaches to predict "*in silico*" the dynamics of proteins: the more classical, atomistic view, and the coarse-grained, low resolution view.

2.1 Molecular and Essential Dynamics:

2.1.1 Molecular Dynamics: describing motions from Newton Laws

An historical overview: Laplace's vision of Newton Mechanics

Molecular Dynamics (MD) is a set of computational techniques to simulate the behavior of atoms and molecules based on the laws of physics. It appeared in the late 50s from the theoretical physics, studying interactions between hard spheres (Alder & Wainbright, 1957). The approach was later extended to study simple liquids such as noble gases (Rahman, 1964), and soon was applied to material sciences. Yet in the 70s, the first realistic simulation with molecular interactions was that of liquid water (Stillinger & Rahman, 1974), whereas the first dynamics of a protein was the small BTPI (*Bovine Trypsin Inhibitor)* which was simulated for barely 3*ps* in gas phase (J A McCammon et al., 1977) (see *Figure 2. 1).*

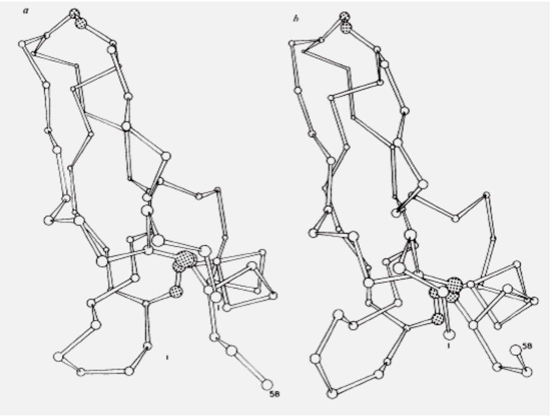


Figure 2. 1 The first MD simulation of a protein by McCammon, Gelin and Karplus (1977).
C-alpha backbone and disulphide bonds of BPTI, before (left) and after (right) 3.2 picoseconds.

The main theoretical basis of MD is Boltzmann's statistical mechanics - for this reason, has been also referred as *"statistical mechanics by numbers"* - but its formulation requires Newton's mechanics. In MD simulations, thermal motion drives a random walk for each atom, whereas bonds and steric repulsions restrict the degrees of freedom to the chemically meaningful region of the Ramachandran plot. The result is an ensemble of conformations (*snapshots*) across the time, which is assumed to be equal to the average of the statistical ensemble under the **ergodic hypothesis**. To satisfy this equality and assure empirical relevance of the results, we must sample a significant region of the phase space. Since the ultimate goal of MD is to predict the future behavior of a system based solely on its prior knowledge by using the rules of Physics, MD has been called *"Laplace's vision of Newtonian Mechanics"*. In the words of the great French mathematician, **Pierre-Simon de Laplace (1749-1827)**, an Omniscient Calculator, only provided with exact knowledge of the present, could predict the entire future:

"We may regard the present state of the universe as the effect of its past and the cause of its future. An intellect which at any given moment knew all of the forces that animate nature and the mutual positions of the beings that compose it, if this intellect were far-reaching enough to subject these data to analysis, could condense into a single formula the movement of the greatest bodies of the universe and that of the lightest atom; for such an intellect nothing could be uncertain and the future just like the past would be present before its eyes."

Laplace, 1820

The limitations of the deterministic dream: Classical and Quantum Methods

During the last century, the Laplacian dream of finding a link between past and future was definitely broken by the Heisenberg's Principle of Uncertainty and the three-body chaos discovered by Poincaré. Therefore, it is not surprising that when Newtonian mechanics is applied to a complex molecular system such as a protein, some problems and limitations emerge. For example, chaotic effects, that are unavoidable in many-body systems and make MD trajectories extremely sensible to initial conditions and thus irreproducible (Braxenthaler, Unger, Auerbach, Given, & Moult, 1997; de Groot, van Aalten, Amadei, & Berendsen, 1996; Elofsson & Nilsson, 1993), or the neglect of electrons in the calculations, which eliminates many subtleties in the description of electrostatic bonds and forces, such as polarization effects or the way hydrogen bonds are modelled in water and secondary structures (Guvench & MacKerell, 2008; Ponder & Case, 2003). Ideally, we should calculate the potential energy of a molecule energy with a quantum Hamiltonian, resulting of all the nuclei and electrons. In practice, the computational cost makes unfeasible this approach to systems such as proteins, and quantum or hybrid QM/MM methods (Senn & Thiel, 2009) are restricted to small systems where charge transfer processes are decisive, such as bond formation, cleavage or polarization. Therefore, the bases of MD are empirical functions of the potential energy (the *force-fields*), only dependent on the nuclei positions, and where electrons are not considered in an explicit manner, as we will see next.

2.1.2 The Basics: Force-field definition and Integration of Motion's Equations

The acting potential energy: molecular force-fields

Empirical Force-Fields intend to describe the potential energy of a given configuration with simple potential functions of the form:

$$E = E_{bonded} + E_{nonbonded}$$

(Equation 2. 1 *Force-field Potential Energy*)

in order to differentiate covalent and non-covalent interactions (*Figure 2. 2*). Only nuclei positions are considered; electrons are taken into account in an implicit manner, through the values of bond constants. For bonded interactions the potential is usually a sum of harmonic and Fourier terms, such as:

$$E_{nonbonded} =$$

$$\sum\nolimits_{bonds} K_s(l-l_0)^2 + \sum\nolimits_{angles} K_b(\theta-\theta_0)^2 + \sum\nolimits_{torsions} \sum\nolimits_{i=1}^{3} \frac{V_i}{2}(1+\cos(i\emptyset-\xi))$$

(Equation 2. 2 *Bonded Terms*)

to model energy of the stretching, bending and torsion motions between atoms, where *l* and *ϑ* are bond length and angle, K_s and K_b are the associated force constants, and V_i, ϕ and ξ are the amplitudes, torsion angle and phase angle associated with Fourier terms. The non-covalent forces are modeled with a sum of a Coulomb potential for electrostatic interactions and Lennard-Jones inverse exponential function of -6 and -12, to account for short range internuclear attraction and repulsion:

$$E_{nonbonded} = \sum\nolimits_{a,b} \frac{Q_a Q_b}{r_{ab}} + \sum\nolimits_{a,b} \left[\left(\frac{C_{ab}}{r_{ab}}\right)^{12} - \left(\frac{D_{ab}}{r_{ab}}\right)^{6}\right]$$

(Equation 2. 3 *Non-bonded terms*)

Where the constants introducing electronic effects are empirical parameters that must be determined from Raman and infrared spectroscopy data, NMR restraints (Fourier terms), liquid phase simulations, lattice energies and crystal structures (Van der Waals) or ab-initio quantum mechanical calculations (Torsional terms) (**parameterization**) (Monticelli & Tieleman, 2013).

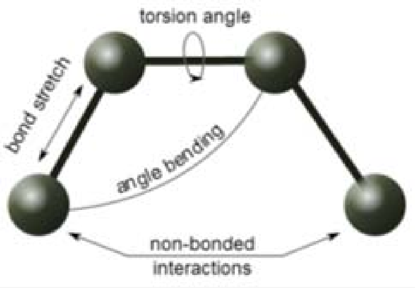


Figure 2. 2. MD Bonded and non-Bonded interactions.

Molecular Dynamics Algorithms and Computational Bottlenecks

In an MD simulation algorithm (see *Figure 2. 3*), we only need an initial configuration of the system, *r* (for example, a set of PDB coordinates) and a force-field, *E(r)*. Since the force acting on each particle is the negative gradient of the potential energy, we can calculate its acceleration:

$$\vec{f}_i = -\frac{\partial E_i}{\partial \vec{r}_i} = m_i \vec{a}_i$$

(Equation 2. 4 *Force*)

From the integration of the acceleration, we can get velocities:

$$\vec{v}_i = \int \vec{a}_i dt$$

(Equation 2. 5 *Velocity*)

And integrating velocities, we can get a new set of positions:

$$\vec{r}_i = \int \vec{v}_i dt$$

(Equation 2. 6 *Position*)

These equations must be integrated numerically, since analytic solutions are only known for the simplest systems. Numerical integration generates a sequence of positions and velocity pairs *{r^n, v^n}* for integers *n* that represent discrete times *t=nΔt* at intervals or *time steps* Δt. The feasible size of the integration step is limited both by the accuracy and the stability of the procedure, constituting a great bottleneck – since the integration step cannot be larger than the fastest motion. The high-frequency modes of motion (bond vibrations) require time steps of the order of 1 fs or less for acceptable resolution, which imply that we need ***one million steps to cover a nanosecond***, a time scale ten orders of magnitude below the slow, large motions of biological significance. On the other hand, larger time steps result in unstable trajectories. Hence, the computational power limits greatly the simulation of biologically relevant dynamics. The most common and simplest family of integrators for biomolecular simulations is the **leapfrog** group, a truncation of a higher-order method developed by Störmer and later adapted by Verlet (Verlet, 1967), which is characterized by exceptional stability over long time when compared to other methods such as **Runge-Kutta**. Since stability is the main limitation to increase the timestep and thus the simulation length, the Verlet algorithm remains still the most popular integrator in MD and all the variations of its basic scheme, from constrained to stochastic dynamics or the different extensions to statistical thermodynamic ensembles (Schlick, 2001).

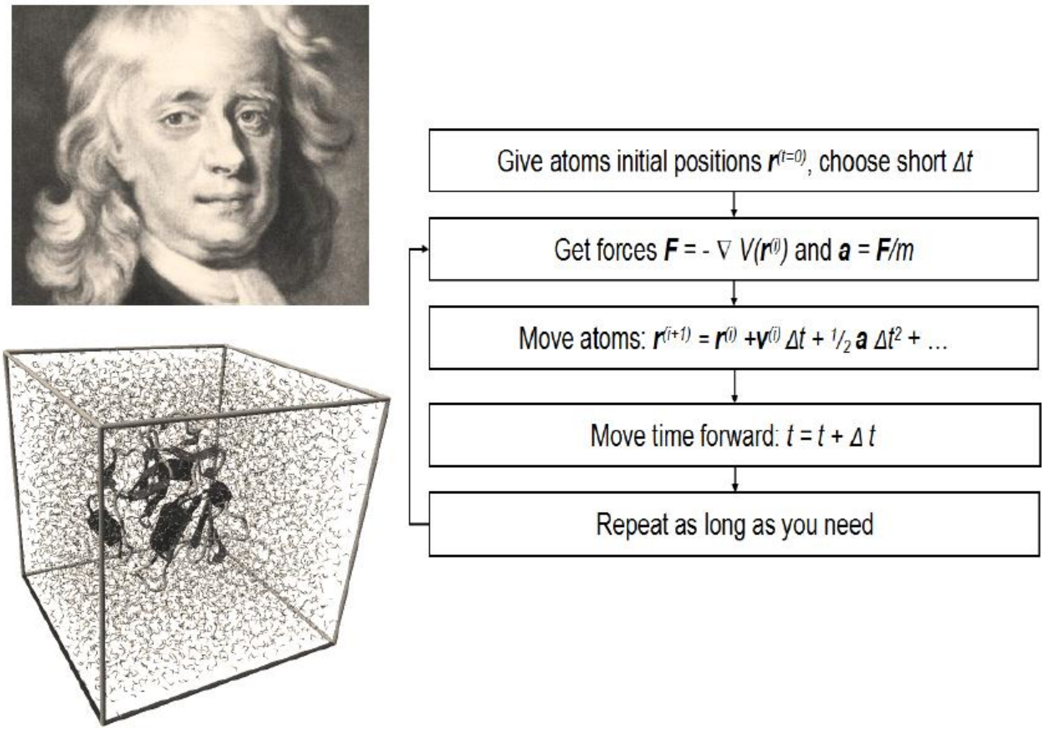

Give atoms initial positions $\boldsymbol{r}^{(t=0)}$, choose short Δt
Get forces $\boldsymbol{F} = -\nabla V(\boldsymbol{r}^{(i)})$ and $\boldsymbol{a} = \boldsymbol{F}/m$
Move atoms: $\boldsymbol{r}^{(i+1)} = \boldsymbol{r}^{(i)} + \boldsymbol{v}^{(i)}\Delta t + {}^{1}/_{2}\,\boldsymbol{a}\,\Delta t^{2} + \ldots$
Move time forward: $t = t + \Delta t$
Repeat as long as you need

Figure 2. 3. Molecular Dynamics Simulation.
The basic algorithm is based on the iterative integration of Newton's Equations of motion (*right*). A typical simulation box is shown (*left*).

2.1.3. Essential Dynamics: a Principal Component Analysis of MD

MD provides a Boltzmann's ensemble of configurations of the protein, from which flexibility descriptors can be derived. Nevertheless, inside the trajectory there is full of noise arising from short-range atomic fluctuations, mixed with the important information on large rearrangements. This complexity makes the analysis of MD simulations hard, rendering it difficult to uncover motions of interest or functional mechanisms. To analyze the trajectories, one can cluster conformations to detect highly sampled regions in conformation space or alternatively, one can employ **Principal Component Analysis (PCA)** to filter the main modes of collective motion from local fluctuations. By a change of orthonormal basis, a complex trajectory is reduced to a lower-dimensional description of the functional motions. This allows enhanced sampling algorithms to search the conformational subspace (Amadei, Linssen, De Groot, Van Aalten, & Berendsen, 1996; De Groot, Amadei, Scheek, Van Nuland, & Berendsen, 1996; De Groot, Amadei, Van Aalten, & Berendsen, 1996; Grubmüller, 1995). The dynamics in the low-dimensional subspace defined by these global modes is called ***"Essential Dynamics"*** (ED) (Amadei, Linssen, & Berendsen, 1993), since these modes are often linked to function, and accordingly, the subspace they define is referred to as ***"essential subspace"***.

To perform a PCA, it is necessary to have an ensemble of conformations from a MD trajectory (or from an experimental source, i.e. NMR, see **Chapter 3**). These multiple conformations must be superposed by a least-squares fit (Kabsch, 1978) to a common reference structure, to filter the internal motions from global rotation and translation. The fitted ensemble is used to build a Cartesian **variance–covariance matrix** of positional fluctuations, C (*Figure 2. 4*):

$$C = \left\langle \left(x(t) - \langle x \rangle\right) \cdot \left(x(t) - \langle x \rangle\right)^T \right\rangle$$

(Equation 2. 7 *Covariance Matrix*)

Where <> denotes an ensemble average; *C* is a symmetric matrix with the variances of each atom displacement as diagonal elements, and as off-diagonal elements, the covariances of the atomic fluctuations for each atom pair relative to their respective averages. Correlated motions give positive covariances, anticorrelated motions negative ones, and non-correlated motions near-zero values. A PCA can be carried out on any subset of atoms, but usually only Cα or backbone atoms are considered. This matrix *C* can be diagonalized by an orthogonal coordinate transformation T, yielding a set of eigenvectors and eigenvalues defining the **Principal Components (PCs)**:

$$C = \Lambda T$$

(Equation 2. 8 *Covariance Matrix Diagonalization*)

where Λ is the diagonal eigenvalue matrix, and T contains, as columns, the eigenvectors. Given a protein of N residues, if only C^{α}'s are taken into account, C is a 3N × 3N matrix. Six eigenvalues must be zero, corresponding to the eigenvectors that describe the three rotational and translational modes along the three axis of the Cartesian space. Then, with at least 3N configurations to build the matrix *C*, 3N-6 eigenvectors with nonzero eigenvalues will be obtained. The eigenvectors e_i indicate the directions of the collective modes in Cartesian space (*Figure 2. 4*), and the corresponding set of eigenvalues, λ_i, describe their mean square fluctuation (the contribution of each component to the total fluctuation); they are sorted according to decreasing order of variance (in units of length squared, Å²). Though there is no harmonic assumption in ED, these largest-amplitude modes match to a high degree the NMA slowest, low-frequency motions, as we will discuss below.

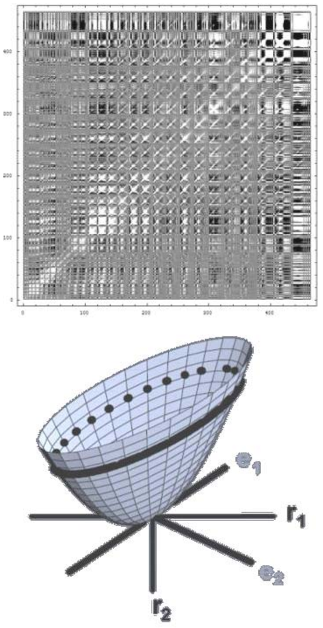
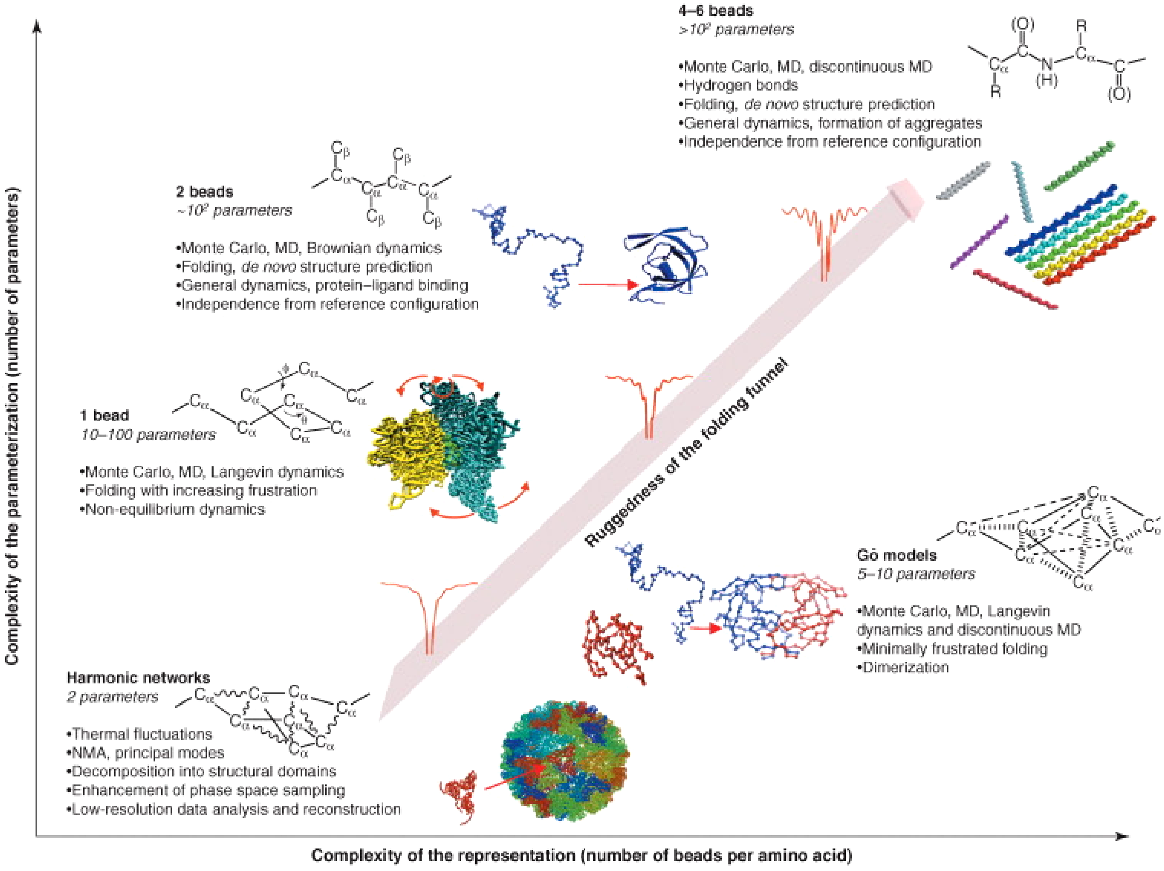


Figure 2. 4. Matrix of variance-covariance from MD and PCs. The diagonalization of the covariance matrix (*above*) yields eigenvectors e_i that represent the preferred equilibrium directions (*below*)

2.2 Coarse-graining: simplifying forces and structures

Coarse-graining is seen by many just as the cheap alternative to costly MD simulations. Technically speaking, it is a process of renormalizing interactions into a new representation with a lower overall dimensionality (Orozco et al., 2011; Saunders & Voth, 2013; Tozzini, 2005). In the case of proteins, coarse-graining usually implies: a) the compression of a series of atoms into pseudoparticles, b) the simplification in the representation of the solvent that can be neglected, simulated as a continuum or also compressed into pseudoparticles and, c) a simplification in the potentials (Ha-Duong, Basdevant, & Borgis, 2009; Rzepiela, Louhivuori, Peter, & Marrink, 2011; Yesylevskyy, Schäfer, Sengupta, & Marrink, 2010). As the number of beads in the united-atom representation of the structural elements decreases, the simulation is faster and the modeled system can be larger (see *Figure 2. 5*).


4–6 beads
>10² parameters
•Monte Carlo, MD, discontinuous MD
•Hydrogen bonds
•Folding, de novo structure prediction
•General dynamics, formation of aggregates
•Independence from reference configuration
2 beads
~10² parameters
•Monte Carlo, MD, Brownian dynamics
•Folding, de novo structure prediction
•General dynamics, protein–ligand binding
•Independence from reference configuration
1 bead
10–100 parameters
•Monte Carlo, MD, Langevin dynamics
•Folding with increasing frustration
•Non-equilibrium dynamics
Ruggedness of the folding funnel
Gō models
5–10 parameters
•Monte Carlo, MD, Langevin dynamics and discontinuous MD
•Minimally frustrated folding
•Dimerization
Harmonic networks
2 parameters
•Thermal fluctuations
•NMA, principal modes
•Decomposition into structural domains
•Enhancement of phase space sampling
•Low-resolution data analysis and reconstruction
Complexity of the parameterization (number of parameters)
Complexity of the representation (number of beads per amino acid)

Current Opinion in Structural Biology

Figure 2. 5. Coarse-grained models classified by complexity of the representation and the parameterization.

For each class of model, the following aspects are reported: schematic representation of the model, indicative number of parameters, methods of solution, main characteristics and applications. Sample applications are also illustrated with representative pictures (prepared using crystallographic coordinates from the PDB [codes 1hhp, 1cwp, 1mwr, 486d]) intended to show the size of system that can be studied and the kind of study that can be done. The location of the models in the *x-y* plane is intended to qualitatively illustrate their complexity, which increases following the direction of the arrows (see (Tozzini, 2005))

The most common level of coarse graining for proteins, as mentioned in **Chapter 1**, implies the representation of every residue by a single particle located at the C_α (Atilgan et al., 2001; Navizet, Cailliez, & Lavery, 2004). This useful compression averages out all chemical properties, and therefore, requires massive re-calibration of the potential functions (the *force-fields*) to include more specific interactions into fewer variables, or to use information-based potentials. Furthermore, CG models are usually parameterized based on a single reference configuration and as a consequence, the dynamics they reproduce is strongly biased towards it. As the graining becomes 'coarser', the parameterization of *force-fields* that are both accurate and transferable becomes increasingly difficult, with different degrees of independence from the reference configuration. Refinements of CG models consist on using additional particles to mimic side chains or backbone atoms (Sacquin-Mora & Lavery, 2006; Zacharias, 2003) or on applying more sophisticated physical potentials (Pasi, Lavery, & Ceres, 2012), which are calculated by comparisons with different sources of flexibility data such as crystallographic B-factors, NMR observables, MD simulations, etc (see examples in (Eyal, Chennubhotla, Yang, & Bahar, 2007; Jernigan, 2007; Kondrashov, Cui, & Phillips, 2006; Kondrashov, Van Wynsberghe, Bannen, Cui, & Phillips, 2007; L. Yang, Song, Carriquiry, & Jernigan, 2008; L. Yang, Song, & Jernigan, 2009; L.-W. Yang et al., 2007)). Next we will review briefly the theory behind the potentials and sampling techniques usually used in coarse-grained methods, to focus later in NMA, which will be explored from many points of view throughout this thesis.

Coarse-grained potentials and sampling techniques

The approach to coarse-grain a biomolecular structure and model its dynamics requires two steps: first, the potential energy of the structure must be described by a simple function, and then, this function must be used to sample the conformational space, a step which consumes the greatest computational power. In general, three kinds of simplified potentials are used in CG simulations:

- **Go-like potentials,** which are in the basis of information-based potentials and are often used in conjunction with Cα coarse-graining (Nobuhiro Go, Noguti, & Nishikawa, 1983). They consider that two residues in contact in the three dimensional structure of the protein have a favorable interaction, while if they are not in contact such interaction is irrelevant. Despite its extreme simplicity, Go-potentials, which assure the principle of minimal frustration, have been successfully applied to study protein folding and coupled to ***Langevin dynamics*** (see *below*), have been used to analyze experimental measures on folding and unfolding.

- **Harmonic potentials:** an evolution of Go-like potentials that widely used for the study of the "near-equilibrium" dynamics of proteins when implemented in sampling techniques derived from normal mode analysis (NMA). Inspired by Flory's networks (Flory, Gordon, & McCrum, 1976) and Rouse *bead-and-springs* models (Rouse, 1953), their basic assumption is that a protein behaves as an ***Elastic Network Model*** **(ENM)** (Tirion, 1996), where C_{α}'s act as nodes which are connected by harmonic springs. The refinement of elastic network harmonic potentials based on protein flexibility data will be the focus of **Chapter 3**.

- **Flat potentials:** discontinuous flat potentials (i.e. stepwise potentials) are based on the concept that the continuum physical potentials can be approximated as a series of discontinuous potentials defined by square wells. The simplest flat square potential is that describing hardcore spheres undergoing elastic collisions, which are then defined by an interaction potential with an infinite step at the distance corresponding to the sum of the radii of two particles, allowing the treatment of trajectories within the ballistic regime. Stepwise potentials have been widely used in the last years in ***discrete Molecular Dynamics (dMD)*** (Proctor, Ding, & Dokholyan, 2011; Shirvanyants, Ding, Tsao, Ramachandran, & Dokholyan, 2012) discussed below, and can be adapted to work with Go-like and pseudo-physical or physical potentials. These potentials can reproduce not only near-equilibrium dynamics, but also local motions like those happening during protein-protein interactions or folding.

- **Physical or pseudo-physical potentials:** these force-fields try to maintain a physical foundation while reducing the degrees of freedom of the system. The most popular example is the MARTINI force-field developed by Marrink et al. (Marrink & Tieleman, 2013; Periole & Marrink, 2013), in which four heavy atoms are represented by a single bead annotated in four types (polar, nonpolar, apolar and charged). The interactions between beads are represented by a physical force-field containing both "bonded" and "non-bonded" terms, very similar to atomistic force fields and that works using the same molecular dynamics algorithms, particularly the GROMACS simulation package. The coarse-graining not only decreases dramatically the degrees of freedom in the system, but also allows the use of large integration steps increasing the time window accessible to simulations.

Irrespective of the nature of the Hamiltonian used to model the dependence of the energy on the protein conformation, the study of flexibility requires the use of sampling techniques. Among the most widely used are the following:

- **Normal Mode Analysis (NMA).** NMA is a technique employed since the 50s for the assignment of vibrational spectra in infrared, Raman or inelastic neutron scattering spectroscopy (Herzberg, 1945). In the last years, it has been established as a common computational tool for the analysis of near-equilibrium protein motions. Instead of numerically solving Newton's equations, NMA assumes the harmonicity of the system (small movements around an equilibrium configuration) and thus allows the computation of a unique analytical solution (i.e. a series of *normal modes*) by expansion of the potential function in a Taylor series (as we will see in next section). Diagonalization of the mass weighted Hessian matrix yields a series of eigenvectors (v_i) and their corresponding eigenvalues (given as frequencies, λ_i), which define the *normal modes*, i.e. the lowest energy movements of the system.

  NMA can be applied in conjunction with any continuum and differentiable potential function, but in general, *CG*-potential functions, such as the *ENMs*, are preferred to *all-atom* potentials, because they increase the speed of the computation and eliminate the problems derived from the initial energy minimization. As we will discuss throughout this thesis, ENM-NMA methods describe extremely well large biologically relevant movements (Navizet, Lavery, & Jernigan, 2004; F Tama & Sanejouand, 2001; Zheng & Doniach, 2003), and are able to reproduce with reasonable accuracy experimental B-factors (Kondrashov et al., 2006; Kundu, Melton, Sorensen, & Phillips, 2002), as well as the pattern of flexibility detected in NMR experiments (L.-W. Yang et al., 2007; L.-W. Yang, Eyal, Bahar, & Kitao, 2009a) or MD simulations (Orellana et al., 2010; Romo & Grossfield, 2011; Rueda, Chacón, & Orozco, 2007). Normal modes are also used to improve fitting of protein structures to electron density maps (Lopéz-Blanco & Chacón, n.d.; Suhre, Navaza, & Sanejouand, 2006; Florence Tama, Miyashita, & Brooks, 2004), to introduce flexibility in small molecule or protein-protein docking (Court, 2009; Dobbins, Lesk, & Sternberg, 2008; Lindahl & Delarue, 2005; Rueda, Bottegoni, & Abagyan, 2009), or to guide sampling in large conformational transitions (Sfriso et al., 2012), among others. Although NMA is usually performed in Cartesian space, the implementation in internal coordinates such as the dihedral torsional space can further reduce the degrees of freedom of the system and increase computational efficiency (see for example recent examples (Dos Santos, Klett, Méndez, & Bastolla, 2013; Mendez & Bastolla, 2010) and an application of (Lopéz-Blanco, Garzón, & Chacón, 2011) in **Chapter 5**)

- **Monte Carlo (MC).** In the *Metropolis* procedure (Metropolis, Rosenbluth, Rosenbluth, & Teller, 1953), the conformational landscape is sampled by perturbing randomly an initial configuration (X_0) to generate a trial configuration, which is accepted ($X_1 = X_o^{'}$) if its potential energy is smaller than the starting one ($E(X_0^{'}) \leq E(X_0)$); otherwise, the acceptance or not ($X_1 = X_0$) depends on a probability function, usually, a Boltzmann distribution for a given temperature. This process is repeated millions of times to allow for a proper Boltzmann's sampling of all the degrees of freedom. As in NMA, in spite of the fact that the lost of time coordinate in the simulation poses a major problem, for example, in non-equilibrium processes, MC has been extremely useful to model protein folding (Hansmann & Okamoto, 1999; Jorgensen & Tirado-rives, 1996). Strategies to couple Monte Carlo with ENM-NMA have been suggested (Rueda, Chacón, et al., 2007) to activate protein movements as displacements along the normal modes.

- **Langevin and Brownian dynamics (BD)** assume that the motion of a particle (of mass *m*) in a fluid is due to the molecular-thermal agitation of the surrounding solvent (which lead to random collisions on the particle, $\vec{\xi}$) and to a dispersive force accounting for the viscous resistance the particle feels on going through the fluid ($-\gamma \vec{v}$) at velocity $\vec{v}$. Implementation of these equations with different type of pseudo-physical coarse-graining is straightforward using standard atomistic MD codes. Specific methods (Carrillo, Laughton, & Orozco, 2012; Emperador, Carrillo, Rueda, & Orozco, 2008) have been developed to deal with simpler pseudoharmonic potentials, such as the ENMs typically implemented in NMA samplings (see **Chapter 5**).

- **Discrete Molecular dynamics (dMD).** This MD-like technique is based on the assumption of a *ballistic regime*, i.e. that the particles move at constant velocity in flat well potentials (see above), which avoid the need of integration of Newton's equations of motion since the trajectory progresses from collision to collision, irrespective of the collision time (Emperador et al., 2008). Despite its simplicity, dMD provides reasonable approximations to the real dynamics of proteins, and it is especially useful in systems with very slow dynamics, for example, diffusion and protein aggregation processes. Discrete-MD can be coupled to any potential function that can be represented by multiple flat well potentials. Together with normal modes directed sampling, dMD allows for the fast exploration of conformational transition pathways (see **Chapter 5**)

A more detailed description of coarse-grained potential energy functions and sampling algorithms can be found in (Orozco et al., 2011).

2.3 Coarse-grained Normal Mode Analysis: Elastic Network Models (ENMs)

As mentioned, **Normal Mode Analysis** (**NMA**) (Case, 1994) is widely used as a simulation method to predict the large-scale motions in biomolecules. Atoms in a molecule are considered coupled by harmonic oscillators, which fluctuate around their equilibrium positions. Therefore, NMA is just an extension of the classical problem of the two-coupled harmonic oscillators (*Figure 2. 6*), to an N-coupled system with 3 degrees of freedom. It is based on the assumption that the conformational energy surface at an energy minimum can be approximated by a Taylor series truncated at the second order term (the so-called **harmonic hypothesis**). It can use any force-field included those developed for MD; however, most applications imply the use of simpler potentials. The pioneering work of Tirion (Tirion, 1996), that proposed a very simple all-atom NMA with uniform harmonic potentials, inspired the concept of an **Elastic Network (EN)** to model equilibrium dynamics some years ago (I Bahar et al., 1997; Haliloglu, Bahar, & Erman, 1997). As the name suggests, the protein is modeled as a simple network of atoms or residues connected by elastic springs. There is no conceptual difference between these EN-NMA and standard atomistic NMA other than the simpler force field, and the very convenient assumption that the reference experimental structure is a minimum in the potential energy function, as we will discuss in the following sections.

2.3.1 Normal Mode Analysis and the Limits of the Harmonic Hypothesis

Normal Mode Analysis Formulation

Given $r = \{r_i = (r_{i,1}, r_{i,2}, r_{i,3})^T : i = 1, \ldots, n\}$, the 3N-dimensional set of vectors representing N atom coordinates in the cartesian space, and a potential energy function U(r) as defined by a force-field. First, we define a stable conformation, R_{min}, representing a local minimum in the potential energy surface, by means of a minimization algorithm. We assume that the potential energy follows a quadratic function ($U \sim X^{2}$) around this equilibrium point. If we expand the function U(r) in a second-order Taylor series, ignoring third and higher-order derivatives, the expansion is truncated at the quadratic level. Thus, the energy surface is approximated by a parabola characterized by the second derivatives evaluated at the equilibrium conformation [11]:

$$U(r) \approx U(R_{\min}) + \frac{1}{2}\Delta r^T H \Delta r, \quad H = \nabla^2 U(R_{\min}),$$ (Equation 2. 9 *Harmonic Hypothesis*)

[1] * We are assuming an energy minima, so that $U(R_{min})=0$

where Δ*r* = *r* – *R*$_{\mathrm{min}}$, and H is the Hessian matrix of the system, a *3Nx3N* symmetric and defined positive matrix, whose elements K_{ij} , the hookean constants associated to the harmonic potential, are the second derivatives of the potential energy U with respect to the mass-weighted atomic coordinates, *dr*:

$$K_{ij} = \left[\frac{\partial^2 U}{\partial r_i \partial r_j} \right]_{r=R_{\mathrm{mir}}}$$

(Equation 2. 10 *Hessian Matrix*)

The matrix H is called the stiffness matrix in classical mechanics and describes the shape of the potential surface. To calculate the vibrational frequencies for this system and the directions of the corresponding motions, we must solve the eigenvalue problem:

$$H.e_i = \lambda_i e_i, \quad i = 1,\ldots N$$

(Equation 2. 11 *Hessian Diagonalization*)

The normal mode vectors are the eigenvectors, e$_i$, of the hessian matrix H, and their associated eigenvalues, λ$_i$, represent the shape of the potential well in the direction of these modes. After removing the six eigenvalues equal to zero for the translation and rotation modes, a non-lineal molecule of N atoms will have 3N-6 normal modes. Displacements in these directions are independent, and hence, the normal modes form an orthogonal basis. In NMA the modes are sorted by frequency, being the lowest frequency modes those with the greatest fluctuation (N Go, 1990), i.e. those explaining most of structural variance.

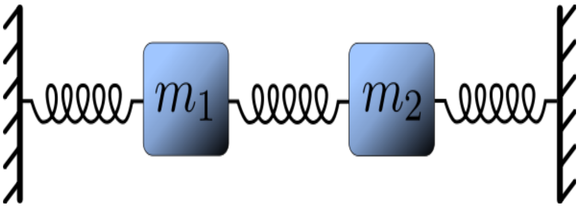


Figure 2. 6. Two coupled harmonic oscillators. The NMA approach is a generalization of this problem to the N-coupled case in the three dimensional cartesian space. The protein is considered a set of N coupled harmonic oscillators, so that the normal modes of vibration can be calculated straightforward from the diagonalization of the matrix containing hookean constants.

The limits of the harmonic approximation: a rugged energy landscape

The harmonic hypothesis exposed in the former section assumes that:

- ✓ the reference structure is an energy minimum,
- ✓ no other minima are populated, and
- ✓ all displacements from the reference structure are harmonic

One has to be aware of these strong assumptions and its limitations. The fact that the harmonic potential is defined around a particular point, and that only small harmonic departures are allowed, apparently makes it difficult to trace transitions from one local minimum to another. Energy landscapes of proteins were first introduced to explain how proteins overcome the **Levinthal paradox** (Levinthal, 1969), but became also useful to understand protein dynamics. The energy landscape of an N-atoms protein is a theoretical construct in 3N-6 dimensions, where N is the number of atoms in the protein and its hydration shell and each multidimensional point describes a structural substate of the protein. The complex landscape is organized hierarchically, with multiple barriers and valleys within valleys; proteins diffuse between substates crossing energy barriers of various heights (see *Figure 2. 7*) which correspond to motions in different timescales (Henzler-Wildman & Kern, 2007). These internal motions have different amplitudes and frequencies: from bond vibrations at the femto- to pico-seconds time scale, or side-chain rotations at nanoseconds, to motions of flexible termini and loops, large concerted domain motions or conformational changes upon ligand binding in larger micro- to mili-seconds time scale. While on short timescales the dynamics of proteins are dominated by fluctuations within a minimum, on longer timescales, the major modes of collective motion are mainly anharmonic transitions between minima, in principle beyond the normal modes approach.

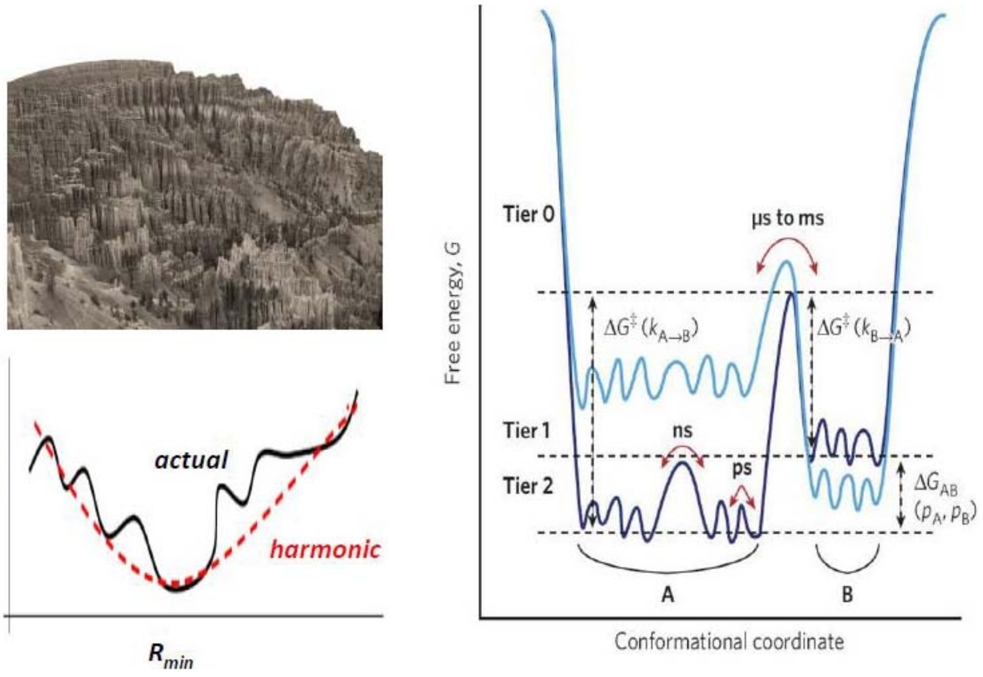

U(R)
actual
harmonic
R_{min}
Free energy, G
Tier 0
μs to ms
$\Delta G^{\ddagger}\ (k_{A\rightarrow B})$
$\Delta G^{\ddagger}\ (k_{B\rightarrow A})$
Tier 1
ns
ps
Tier 2
ΔG_{AB}
(p_A, p_B)
A
B
Conformational coordinate

Figure 2. 7. The rugged energy landscape of proteins and the smooth harmonic approach.
The rugged energy landscape of protein motions (*upper left*) can be approached by an ideal harmonic parabola (*lower left*). A hierarchy of energies and timescales of the landscape defines motions (*right*)*

*Published in: Katherine Henzler-Wildman & Dorothee Kern; *Nature.* **2007,** 450, 964-972. DOI: 10.1038/nature06522 Copyright © 2014 Nature Publishing Group

Nevertheless, though NMA ignores the "*rugged*" nature of energy landscapes, normal modes are often more correct that one might expect. Despite the energy surface contains many local minima, proteins tend to behave as this surface was harmonic. In principle, conformational changes involve transitions between different potential wells out of the harmonic regime, but from dynamical systems theory, it is known that NMs are very robust under small perturbations. Many studies that compare NMA modes with functional transitions derived from experimental data confirm that the normal modes with the largest fluctuation (lowest frequency modes) are indeed both biologically and functionally relevant (Petrone & Pande, 2006). Therefore, most conformation changes in proteins can be strikingly well described by a small number of modes, which predict with great accuracy functional large-scale changes such as domain or hinge-bending motions (Gerstein & Krebs, 1998; Krebs et al., 2002; F Tama & Sanejouand, 2001).

2.3.2 Elastic network-Normal Mode Analysis

Elastic Network Models (ENMs) represent a further step in the simplification embodied by NMA. All ENM assume that protein flexibility is due to harmonic deformations around a reference structure. The first ENM, proposed by Tirion (Tirion, 1996), was an all-atom model with a simple pairwise Hookean potential. In Tirion's model, the native structure is defined as a minimum, and the detailed atomic force-fields are replaced by a simple harmonic potential, with a uniform constant to all interactions within a cutoff. Tirion showed that this minimal model reproduced both the low-frequency modes and the Cα fluctuations accurately. Later, the **Gaussian Network Model** (**GNM**) (I Bahar et al., 1997) introduced the coarse-graining of the protein structure, reduced to the Cα backbone. Mathematically, though GNM is defined as a coarse-grained ENM reduced to one coordinate per atom, it is very different from a physical point of view. As the ENMs, it predicts the atomic fluctuations and, via the correlations, some dynamical information, but cannot provide information about the directions of motions. Due to its insensitivity to the directions, GNM assigns a non-zero energetic cost to global rotations, therefore violating the principle of invariance under rotation. On the contrary, the ENMs are physical models describing small-amplitude harmonic departures from a stable conformation. The **Anisotropic network model** (**ANM**) (Atilgan et al., 2001), that we will use in this work, is often presented as an extension of the GNM to the 3-D space to consider directionality. However, it is nothing else but an ENM for the C^{α} atoms *(Figure 2. 8*), as we will explain in the next section.

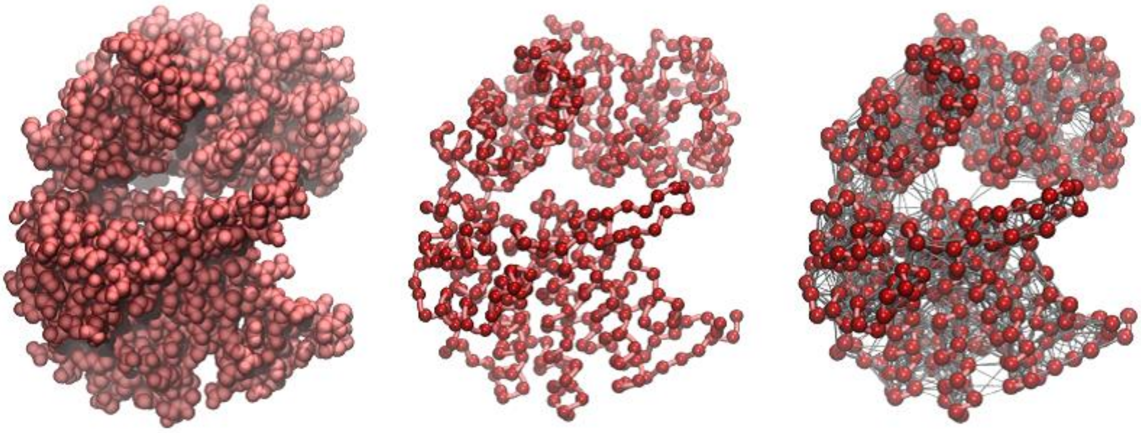


Figure 2. 8. Coarse-graining of the 3-D structure as an Elastic Network.
The protein is represented by the C-alpha carbon trace (left), where each residue or node is connected to other residues with Hookean springs (right) modelling intramolecular interactions.

Elastic Network Model Formulation

In the original Tirion's form, nodes in the network are identified by the positions of all atoms, and all nodes within a cutoff distance are connected with a uniform force-constant. The network topology is then described by a **Kirchhoff matrix Γ** of inter-residue contacts where the *ij*-th element is equal to -1 if nodes i and j are within the cutoff distance r_c, and zero otherwise, and the diagonal elements (ii-*th*) are equal to residue connectivity:

$$\Gamma_{ij} = \begin{cases} -1 & if\ r_{ij} \le r_c \\ 0 & if\ r_{ij} > r_c \end{cases} \qquad \Gamma_{ii} = -\sum_{j|j\neq i}^{N} \Gamma_{ij}$$

(Equation 2. 12 *Kirchhoff Matrix*)

The potential energies between each residue pair i-j are given by:

$$U(r) = \frac{\gamma}{2}\Gamma_{ij}\left(r_{ij} - r_{ij}^{0}\right)^2$$

(Equation 2. 13 *Pair Energies*)

And the overall potential of the system is given by a sum of these harmonic potentials:

$$U(r) = \frac{\gamma}{2}\sum_{j/j\neq i}\Gamma_{ij}\left(r_{ij} - r_{ij}^{0}\right)^2$$

(Equation 2. 14 *Hamiltonian*)

where r_{ij} and r_{ij}^0 are the instantaneous and equilibrium distances between nodes *i* and *j*, Γ_{ij} is the ij-th element of the Kirchhoff matrix, and γ is the spring, or force constant for the elastic bond between the atoms and is the same for all atoms pairs; this summation is only performed over atoms less than the cut-off distance r_c.

For the sake of simplicity, the product of the topology matrix and the force constant can be expressed as a single parameter:

$$K_{ij} = \gamma \cdot \Gamma_{ij}$$ (Equation 2. 15 *Stiffness Matrix*)

Then, the molecular Hamiltonian describing the elastic energy to displace a protein from its equilibrium conformation can be expressed as:

$$U(r) = \sum_{j/j \neq i} K_{ij} \left(r_{ij} - r_{ij}^0 \right)^2$$ (Equation 2. 16 *Elastic Energy*)

The energy function in *Eq.2.16* is the most widely used, although others such as the GNM are possible; we will not discuss them. When nodes in the network are identified by the positions of C^{α} atoms, the energy function corresponds to the **ANM**. Actually, this functional can be implemented into Monte-Carlo or dynamics algorithms (Emperador et al., 2008) to obtain ensembles of accessible configurations. Within the NMA described above it is used to build the Hessian matrix (***H***); for a protein network of N nodes (residues), the Hessian matrix ***H*** is a *3N x 3N* matrix consisting of submatrices ***H$_{ij}$***. The *N x N* out-diagonal super elements, ***H$_{ij}$*** ***(i≠j)*** are found from the second derivatives of ***V*** with respect to node positions:

$$H_{ij} = \frac{K_{ij}}{(r_{ij}^0)^2} \begin{bmatrix} X_{ij}X_{ij} & X_{ij}Y_{ij} & X_{ij}Z_{ij} \\ Y_{ij}X_{ij} & Y_{ij}Y_{ij} & Y_{ij}Z_{ij} \\ Z_{ij}X_{ij} & Z_{ij}Y_{ij} & Z_{ij}Z_{ij} \end{bmatrix}$$ (Equation 2. 17 *Hessian Out-diagonal elements*)

Where X_{ij}, Y_{ij} and Z_{ij} are the components of the cartesian equilibrium distance vector R_{ij}^0. The diagonal submatrices of H are defined as follows:

$$H_{ii} = -\sum_{j,j \neq i} H_{ij}$$ (Equation 2. 18 *Hessian Diagonal elements*)

Once the hessian has been calculated, the procedure is the same as for standard NMA: diagonalization of the Hessian matrix of Force Constants (also called the *stiffness matrix*) yields the eigenvectors representing the principal modes, and their associated eigenvalues (in energy/frequency units), that indicate their amplitude or vibrational frequencies (*Figure 2. 9*).

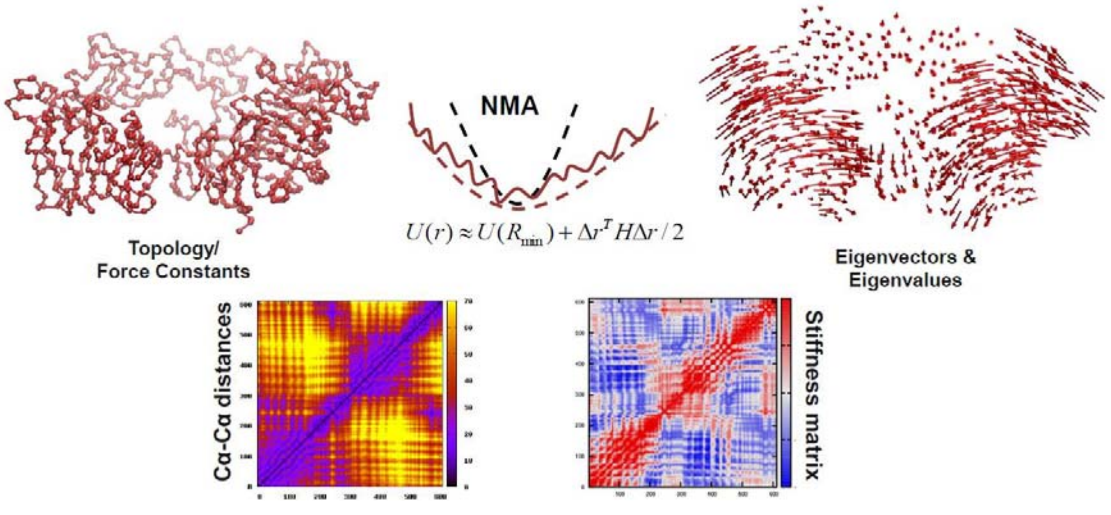


Figure 2. 9. Normal Mode Analysis of an Elastic Network Model.
The protein is represented by the C$^{\alpha}$ trace (*left*), and based on the inter-residue distances a stiffness matrix of spring constants is build and assuming the harmonic hypothesis (*center*) diagonalized to obtain eigenvectors and eigenvalues representing the molecular motions (*right*).

Computational and conceptual simplicity

The atomistic force-field based NMA described in the first section is often referred as "*standard*" NMA to be distinguished from the coarse-grained EN-NMA. As we have seen, standard NMA requires three calculation steps:

1. **Minimization of the potential energy** as a function of the atomic coordinates;
2. **Calculation of the second derivatives** of the potential energy, the Hessian matrix
3. **Diagonalization of the Hessian** matrix yielding normal modes

Depending on the size of the molecule, each one of these steps can have a significant computational cost; the bottlenecks are often the first and final steps: energy minimization demands CPU time, whereas numerical diagonalization needs, in addition, memory. This explains the current popularity of the ENMs, that are still NMA, but with a dramatically simplified force-field. The standard NMA is performed on *all atoms* as the force field requires, but the ENM can be restricted to a subset; in addition, the reference structure is assumed to be a minimum. Consequently, there are two advantages: 1) **there is no need of energy minimization**, since the distances of all springs are taken at their equilibrium length, and 2) **the diagonalization is easier** because only Cα are considered, leading to a tenfold reduction in the hessian. The main drawback is that there are two parameters to be set: the γ force constant, and the cut-off distance, r_c; we will discuss them next.

2.3.3 Network topology and weighting in the ENMs

Several attempts have been made to improve the performance of ENMs, and all of them are based on the refinement of the network connectivity and the associated energy functional. In other words, they aim to find an optimal definition for the nodes and for the strength of the simplified interactions that will connect them: *when* and *how* strong nodes are linked *i.e.* the *connection rules* to build the network. Though the original model by Tirion considered all the atoms as nodes, current ENMs reduce typically each residue to the position of the alpha-carbon. However, it is possible to increase the resolution introducing other atoms in the network, such as the beta-carbons to account for the side chains. On the contrary, some models decrease the resolution by grouping atoms into rigid blocks **Rotating/Translating Blocks (RTB)** (Durand, Trinquier, & Sanejouand, 1994; Li & Cui, 2002; Navizet, Lavery, et al., 2004; F Tama, Gadea, Marques, & Sanejouand, 2000). In this thesis we will focus in the most widely used and natural level of coarse-grain, based on the C-alpha carbon network.

Connectivity rules: Cutoff Models versus Continuous Functions

All physical forces in matter decrease with distance: the interactions between a pair of nearby atoms are on average stronger than the interactions between a distant pair. This property implies that the force constant linking two atoms must be inversely proportional to a power of the distance between them. According to this, connectivity rules are mainly based on the physical distance between interacting nodes. The rules range from a discrete cutoff function to decide the status of connectivity, to continuous functions connecting all nodes in the network with a distance-decaying strength (*Figure 2. 10)*. Then, the selection of the spring constant can be further divided into two parts: i) ***which nodes in the network will be connected/interact?*** , and ii) ***which is the magnitude of the force constant assigned to an interaction?***

Tirion's original all-atom model used the simplest possible approach to assign the force constant: a step function with the same spring for all pairs within a cutoff and zero otherwise. This approach has the drawback that small changes in the input conformation can modify significantly the interactions. The choice of the cutoff also introduces a source of arbitrariness, and therefore, other approaches have been developed to replace the discontinuous Hamiltonians by continuum functions based on the scaling down of the force-constants with distance between nodes (K Hinsen, 1998; Kovacs, Chacón, & Abagyan, 2004). Though the continuum approaches avoid the problems intrinsic to the use of an empirical cutoff, the introduction of remote interactions (relevant or not) can introduce noise and extra rigidity in the network, besides increasing the computational cost.

Usually, connectivity rules do not take into account the chemical properties of the residues or the physics of their interactions (which is partially implicit in the structure, since each force has a typical range length), but some authors do have explored different scaling of covalent and non-covalent interactions (Konrad Hinsen, Petrescu, Dellerue, Bellissent-Funel, & Kneller, 2000; Jeong, Jang, & Kim, 2006; Kondrashov et al., 2006). The C^{α}- C^{α} distance as criteria to assign couplings seems most reasonable and simple, because reflects the kind of interactions, dependent on the chemical identity. Nevertheless, there are other parameters to assign springs, such as the packing environment, but we will not discuss them (Chennubhotla & Bahar, 2007; Sen & Jernigan, 2006). Since any decaying function has a horizontal asymptote when the distance goes to infinity, it is also possible to define a threshold distance beyond which the constant approaches to zero – though strictly speaking never reaches it. In the case of an inverse exponential, the arbitrary choice of the exponent plays the same role as the cutoff distance, since it determines the inflexion point of the curve – that would correspond to the discontinuity in a cutoff function. In practice, the important question is not to determine if it is preferable to use a continuous or discrete function, but to find the distance beyond which interactions become irrelevant, and how can be related to known factors such as chain length or packing density. We will try to address this question in the next Chapter of this thesis.

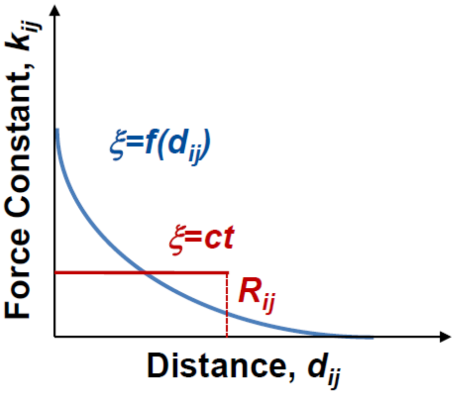


Figure 2. 10. Force Constants: discrete and continuous distance-dependent functions.

A discrete step function (*red*) based on a threshold distance (R_{ij}) and a continuous inverse exponential of the distance (*blue*)

Cartesian distance and sequential distance

The EN models are closely related to the *bead-and-strings* **Rouse chain model for polymers** (Rouse, 1953). However, Rouse chains only connect sequentially adjacent beads, whereas in the ENMs distant pairs in close contact are coupled, in addition to neighbours along the sequence. This modification allows for the relevance of the non-covalent interactions that make the 3-D structure of proteins. By relying exclusively on the Cartesian distance between nodes, ENMs ignore sequential information, and thus are not able to distinguish between residues directly connected (strong covalent bonds) and distant interactions (weak, non-covalent interactions) (see *Figure 2. 11*), which can generate artifacts in the representation of protein flexibility.

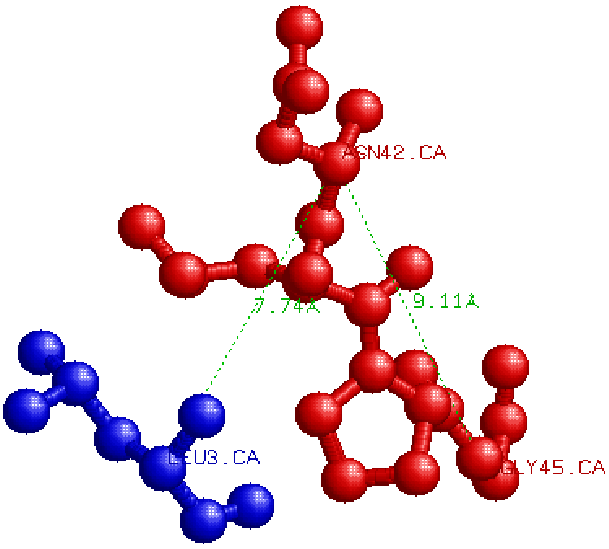


Figure 2. 11. Spring Assignation with Ca-Ca distance.

Example from a training set protein (1sur, 215 residues). Standard cutoff values around 8-9 Å discard the chain coupling between ASN42 and its close neighbor GLY45, but introduce an artefactual bond with the distant, non-chemically interacting residue LEU3. Distance-dependent functions assign a lower force constant to the closer, i→i+3 coupling (ASN42-GLY45), than to the far i→i+41 interaction (LEU3-ASN42).

2.3.4 Experimental and Theoretical Validation of ENMs

The rigorous validation of coarse-grained approaches is not trivial, since experimental data on protein flexibility is scarce and quite indirect. Some works modelled chemical interactions using the magnitude and direction of computational variance as optimizing parameter (Kondrashov et al., 2007; Leioatts, Romo, & Grossfield, 2012), others have relied mostly on comparisons with experimental B-Factors (Riccardi, Cui, & Phillips, 2010; Xia, Tong, & Lu, 2013; L. Yang et al., 2009), but did not assessed other flexibility descriptors such as similarity with principal components. Thus, concern exists on whether a small advance in the quality of the model compensates the increase in model complexity and the need for adjusting more *ad hoc* parameters. Next we briefly discuss some of the experimental and theoretical sources of accurate data and flexibility, and how they compare with ENMs.

X-ray crystallography: B-Factors and multiple conformers

X-ray crystallography is a method to determine the atomic and molecular structure of materials based on the diffraction of a beam of X-rays by a single crystal into many specific directions. The angles and intensities of the diffracted beams are registered in a photographic plate or any suitable detector at different orientations as a diffraction pattern of regularly spaced spots known as *reflections*. Multiple two-dimensional images taken at different rotations allow reconstructing a three-dimensional model of the electron density within the crystal using Fourier transforms. The mean positions of the atoms and their chemical bonds - i.e. an atomic model of the molecule – are determined by iteratively fitting to the electron density map, with a resolution that depends on the size of the crystal or the degree of disorder among other factors. When single crystals of sufficient size are not available, other X-ray methods such as ***Small-Angle X-ray Scattering (SAXS)*** or ***Electron crystallography*** are useful to obtain lower resolution information.

The idea that crystals could be a diffraction media for X-rays was first envisioned by **Max Von Laue**, who realized that electromagnetic radiation of a wavelength comparable to the unit-cell spacing could allow inferring data from atomic structures. Von Laue not only obtained the first diffraction pattern from a copper sulfate crystal but also deduced the law that relates the scattering angles and the size and orientation of the unit-cells, which granted him the Nobel Prize in 1914. The next year the **Braggs'** (father and son) developed the law that relates the observed scattering with reflections, for which they shared the 1915 Nobel Prize.

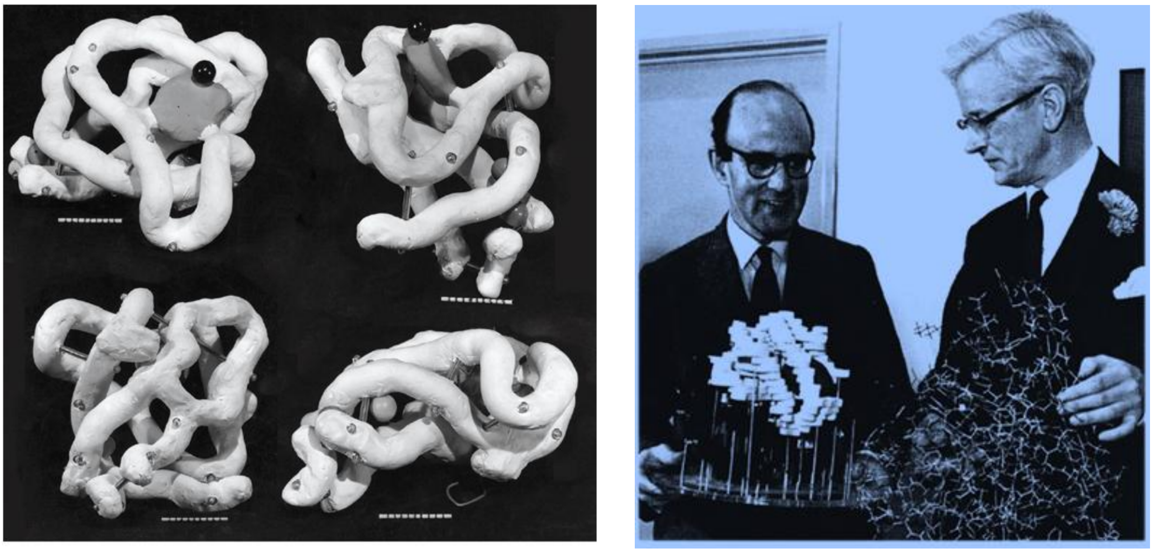


Figure 2. 12. First crystal structures of proteins by Kendrew *et al.* (1958).
Left: The low-resolution structure of myoglobin solved by John Kendrew and colleagues as appeared in the original 1958 *Nature* paper. Polypeptide chains are in white and the grey disc represents the haem group. Note that in spite of the coarse-graining of the structure (with marks on the scale at the astonishing distance of 1 Å apart) the shape is perfectly recognizable. *Right:* Max Perutz (left) holding a manually-built wood model of haemoglobin structure solved at 6-Å resolution, and John Kendrew (right) holding a wire model of myoglobin at 1.4-Å resolution (1962).

The potential of X-ray crystallography for determining the structure of molecules was soon realized and, as the field rapidly evolved from obtaining the structures of simple inorganic crystals (such as table salt, the first atomic-resolution structure to be "solved") to more complex organic molecules such as fatty acids or porphyrins yet in the 20-30s. The crystallography of biomolecules advanced dramatically with the work of **Dorothy Crowfoot Hodgkin**, who solved the structures of cholesterol (1937), penicillin (1946) and vitamin B12 (1956) for which she was awarded the Nobel Prize in 1964. During the 60s, the first low-resolution structures of proteins were obtained: first sperm whale myoglobin, by Sir **John Kendrew**, and later haemoglobin, by **Max Perutz**, for which they shared the 1962 Nobel Prize (*Figure 2. 12*). There are more than 80000 crystal structures deposited in the *Protein Data Bank* (Berman et al., 2000). However overexpressing a protein and obtaining crystals is extremely challenging, and despite advances such as high-throughput techniques or synchrotron radiation, it poises great difficulties for example in the case of very flexible or disordered proteins and specially in membrane proteins – underrepresented in the PDB with less than 100 structures, although they constitute 1/3 of the genes and play key biological roles.

An intrinsic limitation of crystal structures comes from the unnatural environment that represents the ordered lattice, which largely restricts large-scale movements: as discussed in **Chapter 1**, crystal structures are static pictures of an energetically minimized conformer under crystallization conditions, although functional proteins in the cell are best represented by an ensemble of different conformations. Occasionally, a pair of extreme conformers can be captured – such as open/close unbound and bound states (see **Chapter 3**). More often the only information on flexibility is reduced to the **Debye-Waller factors (DWFs) or temperature factors,** used to describe the attenuation of x-ray scattering caused by thermal motion. In protein crystallography, the variability of the atomic positions is described by a symmetric fluctuation tensor composed of six independent elements called *"Anisotropic Displacement Parameters" (ADPs)*. If resolution is not enough, fluctuations are assumed to be isotropic and thus reduced to a single number per atom commonly known as the B-factor, which is related to atomic fluctuations by a simple relation:

$$B_i = \left({}^{8\pi^2}\!/_{3} \right) . \langle (\Delta r_i)^2 \rangle \qquad \text{(Equation 2. 19 Crystallographic Thermal B-Factors)}$$

and measured in units of Å^2. For the majority of structures in the PDB, the DWFs are considered isotropic and reported as B-factors; only recently some high-resolution structures have computed the ADPs. The B-factors are assumed to indicate the relative vibrational motion of different parts of the structure: low B-factors correspond to well-ordered and rigid regions, whereas large B-factors generally belong to flexible parts. The majority of proposed ENM-like models have relied on the prediction of thermal fluctuations for their comparison with B-factors (Eyal, Yang, & Bahar, 2006; Hamacher & McCammon, 2006; Kondrashov et al., 2007). However, the use of B-Factors as the gold standard for ENMs has several problems. ENMs are applied to single proteins, whereas X-ray crystallography is performed on protein crystals. The fluctuations calculated from an ENM represent thermal motions, whereas the mobility described by DWFs is not only due to the spread of the electron density of vibrating atoms but also to static disorder, crystal defects, experimental artifacts, etc. A study considering an ENM for a whole protein crystal (Konrad Hinsen, 2008) showed that thermal fluctuations indeed contribute very few to B factors. Second and more important, both B-factors and ADPs are not experimental observables, but parameters in a theoretical model that is fitted to the experimental diffraction intensities (the real observables). The goal of the refinement process is to obtain a good structure, and they are included because are needed to obtain a good fit. Usually, restraints are imposed on DWFs to reduce the number of independent parameters to fit, for example, using a theoretical model for collective motions from which the ADPs are derived. The standard model for collective motions is the **Translation-Libration-Screw Rotation** or ***TLS model*** (Schomaker and Trueblood, 1968; Winn et al., 2001) describing the protein as rigid subunits; most worrying, low-frequency normal modes are also used to compute the ADPs used to calibrate NMA (Diamond, 1990; Kidera and Go, 1990; Poon et al., 2007). Therefore the use of model fitted parameters to fit another (or even the same) model seems at least, not ideal. Although the main utility of crystallographic *DWFs* in the study of protein flexibility lies thus in their use to evaluate theoretical models for large-amplitude collective motions, they cannot be used in isolation and caution must be taken in their interpretation.

Nuclear Magnetic Resonance: RDCs and Structural Ensembles

Nuclear Magnetic Resonance (NMR) and X-ray crystallography are still the only methods capable of solving the structures of biological macromolecules at atomic resolution. NMR is based on the splitting of energy levels of atomic nuclei by a magnetic field. Atomic nuclei with nonzero spin (i.e. with an uneven number of nucleons and thus a magnetic moment and angular momentum), such as ^{1}H or ^{13}C, can be excited by electromagnetic radiation whose frequency is equivalent to the energy difference between levels and then relax, re-emitting radiation in a similar wavelength. Usually, transitions between these magnetic-induced energy levels involve frequencies in the radio spectra. Although the basic phenomenon of NMR was discovered in the 40s (Bloch, Hansen, & Packard, 1946; Purcell, Torrey, & Pound, 1946), the technique started to be used in chemistry to study metal complexes in solution in the 50s-60s, and only in the 1980s, when powerful enough equipment and techniques where available, was applied to protein structure determination, led by pioneers such as Kurt Wüthrich, who received the Nobel Prize in 2002 (Wüthrich, 2001).

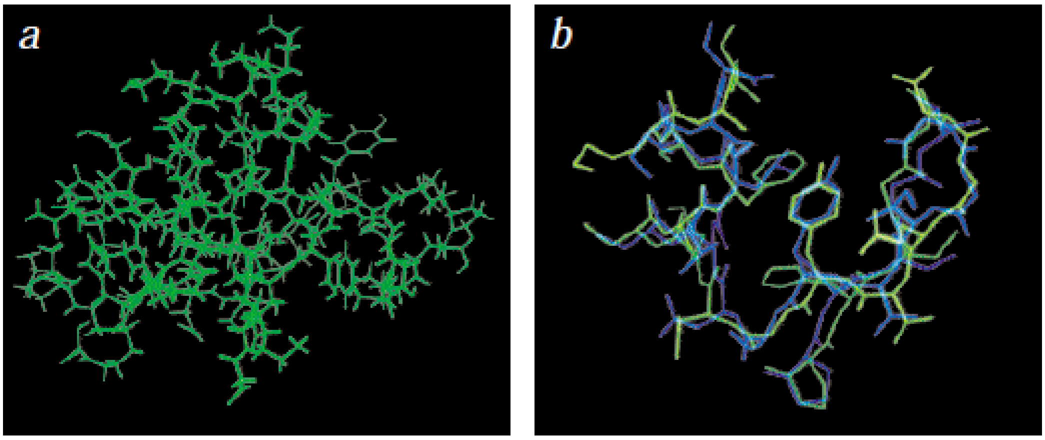


Figure 2. 13. First NMR structures of proteins by Wüthrich *et al.* (1984).
The structure of BUSI (*Left*) was presented informally in 1984, and the reaction was one of disbelief, even receiving accusations of having been modelled after an independent crystallographic study of the homologous protein PSTI. Then, Robert Huber (Nobel Prize in Chemistry, 1988) proposed to settle the matter by independently solving a new protein structure by X-ray crystallography and by NMR. Wüthrich and Huber groups received supply of the α-amylase inhibitor tendamistat from Hoechst, obtaining virtually identical three-dimensional structures (*Right*).

The first protein structure solved by NMR was the small proteinase inhibitor IIA from bull seminal plasma (BUSI II) presented in 1984 (see *Figure 2. 13*) (Williamson, Havel, & Wüthrich, 1985). Nowadays, near 20% of all newly deposited protein structures are solved by NMR spectroscopy.

The limitations of the technique come from its intrinsically low sensitivity and the high complexity of spectra obtained, which usually hampers its application to proteins over 40-60kDa. The sequence-specific assignment of the hundred to several thousand NMR peaks for a protein is possible thanks to multidimensional techniques that simplify spectra and allow determining the experimental restraints that, in combination with computational tools, make possible to elucidate the protein fold. As a counterpart to the size drawback, NMR allows studying proteins in its native aqueous environment which makes the method applicable in principle to intrinsically unstructured or very mobile proteins. Most interesting, it also provides dynamic information on a wide range of timescales, from picoseconds to even days – from reaction kinetics to protein folding or protein-ligand interactions - and thus provides results complementary to the detailed but static information of X-ray crystallography.

NMR experiments range from relatively simple (1D, organic molecules) to quite complex (multidimensional ^{15}N ^{13}C NMR). The NMR method for protein structure determination is based on the **Nuclear Overhauser Effect (NOE)**, which explains spin polarization transfer between two neighboring spin populations via cross-relaxation. Distance restraints originate from NOE measurements, whereas angle restraints are obtained from **J-coupling** (indirect dipole-dipole coupling) between active spin nuclei (such as ^{13}C or ^{15}N) linked by covalent bonds. NOE peak intensities are proportional to r^{-6}, where r is the distance between two spin active nuclei and can only be measured from 1.8 to 6Å. In addition, the signal peak intensity is not exact and needs to be interpreted as a distance range. Usually, secondary structure elements are well defined but hydrophobic core and flexible loops might remain under-determined. The J-coupling depends on the dihedral angles between bonds and thus allows estimating Φ-Ψ values via the *Karplus equation* (Karplus, 1963). Finally, a third class of orientation restraints can be obtained measuring **Residual Dipolar Couplings (RDCs)** from solid-phase or special field oriented NMR, which contains also information on dynamics up to the millisecond range that is inaccessible to other techniques. Initially, NMR ensembles were generated by fitting every structure to all available restraints which led to very rigid ensembles, but recent developments such as **Dynamic Ensemble Refinement (DER)** or **Ensemble Refinement with Orientational Restraints (EROS)**, which introduce besides NOEs, S2-order parameters and RDC data for fitting. As we will see in **Chapter 3**, some recent NMR ensembles contain indeed information on protein flexibility and thus can be used as reference for MD and NMA calculations.

Comparison with Atomistic Simulations

In a previous work, we compared the performance of the classical cutoff approach with the distance-dependent form (Rueda, Chacón, et al., 2007), using as reference the ED predictions in addition to experimental B-factors. The results showed that, though both methods provide a reasonable description of the deformability pattern by MD, the continuous weighting of the constants improved the predictions, approaching the calculated deformations to the ED- calculated modes. However, they fail to describe the pattern of variance found in MD. In all these models, phenomenological force constants are chosen and define an arbitrary energy scale, which correctly describe structural flexibility, but have problems to predict the time amplitude and frequency of the slow motions (Konrad Hinsen et al., 2000).

The striking accuracy of the method: global dynamics does not rely on details

Due to the extreme simplification, it can be questioned the accuracy of EN-NMA versus the standard, atomistic NMA. There is both theoretical and experimental evidence that NMA and ENMs show a high correspondence (Kondrashov et al., 2007). Probably, given the strong assumptions underlying the harmonic hypothesis, the differences between CG and atomistic NMA are irrelevant when both methods are compared with more realistic approaches such as MD. Furthermore, ENM predictions are in good agreement not only with atomistic simulations, but also with experimental data on flexibility (Ivet Bahar & Rader, 2005; J. Ma, 2005; Rueda, Chacón, et al., 2007). Due to the coarse-graining, ENMs have a strong cooperativity and can predict large conformational changes with astonishing accuracy, even outperforming atomistic NMA or ED. Coarse-grain a protein is a drawback when looking at local rearrangements but can even improve the results if we are trying to capture global motions – by adapting the physical model to the scale of the movements we want to describe.

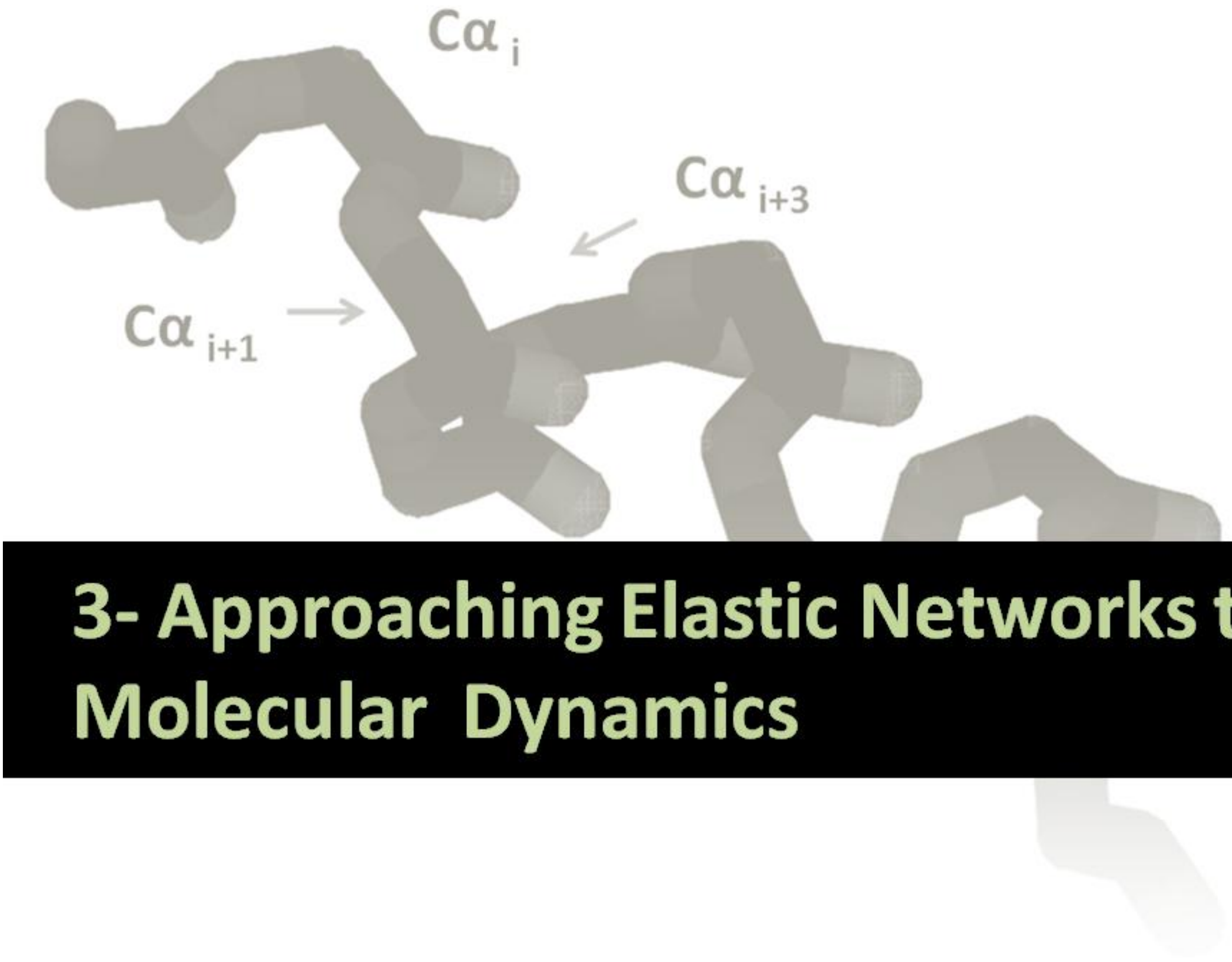
The accuracy of these methods to trace the natural motions, considering their simplicity, is remarkable and is teaching us an important lesson. The fact that large-scale transitions can be reproduced by a network of simple springs, with just one connection per amino acid, shows that these collective motions do not depend on fine details of atomic potentials but rather on the size, shape and general connectivity of the system, as we will explore in the next Chapter.

2.4 Publications from this chapter

Orozco M., Orellana L., Hospital A., Naganathan A.N., Emperador A., Carrillo O. and Gelpí J.L. (2010) ***Coarse Grained Representation of Protein Flexibility. Foundations, successes and shortcomings*. In "Computational Chemistry Methods in Structural Biology".** *Advances in Protein Chemistry and Structural Biology*, vol 80. Ed. C.Christov. Elsevier.

“It is more important to have beauty in one's equations that to have them fit experiment”

Paul Dirac



3- Approaching Elastic Networks to Molecular Dynamics

This page intentionally left blank

Chapter 3 Approaching Elastic Network Models to Molecular Dynamics

In this Chapter, we will perform a thorough comparison of the flexibility derived from ENMs and compare it with ED/MD simulations, in order to optimize the network connectivity rules. In a first phase, we explore the influence of threshold distance and spring strength, to find the connectivity rules for an optimized ANM. We define a nearest-neighbours mixed algorithm, called ED-derived ENM or ED-ENM (Orellana et al., 2010), which is based on the sequential and cartesian distance between Cα pairs. In the second phase, the reference MD is compared with ED-ENM and two classical ENM-NMA implementations: i) ANM in its original implementation with a cutoff function; and ii) ANM with an inverse exponential function. Finally, the new approach is further compared with experimental data from X-ray and NMR ensembles (*Figure 3. 1*).

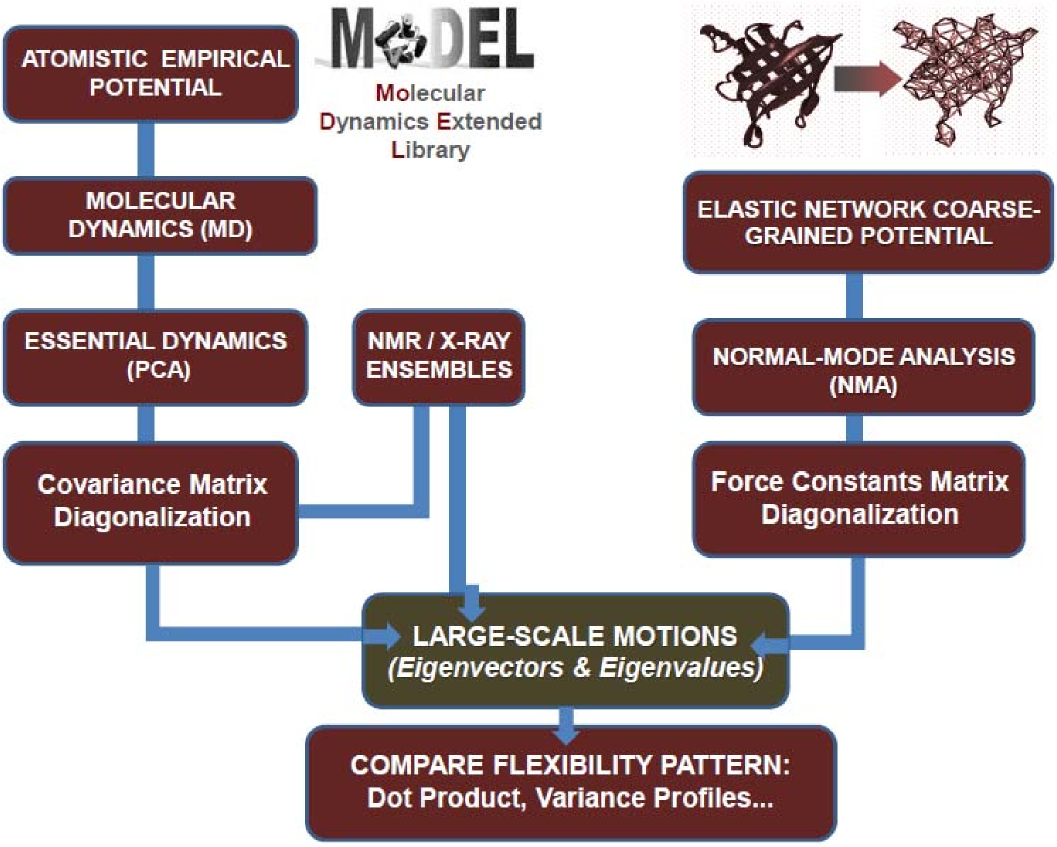

ATOMISTIC EMPIRICAL POTENTIAL
MODEL
Molecular Dynamics Extended Library
MOLECULAR DYNAMICS (MD)
ESSENTIAL DYNAMICS (PCA)
NMR / X-RAY ENSEMBLES
Covariance Matrix Diagonalization
ELASTIC NETWORK COARSE-GRAINED POTENTIAL
NORMAL-MODE ANALYSIS (NMA)
Force Constants Matrix Diagonalization
LARGE-SCALE MOTIONS (Eigenvectors & Eigenvalues)
COMPARE FLEXIBILITY PATTERN: Dot Product, Variance Profiles...

Figure 3. 1 Protocol to compare the flexibility from ENMs with MD and experimental samplings

3.1 The Benchmark Proteins

Our reference set of MD simulations is taken from MoDeL, the largest database of state-of-the-art MD trajectories available (Meyer et al., 2010; Rueda, Ferrer-Costa, et al., 2007b) with a fully atomistic force-field and explicit solvent. This database contains the trajectories for highly representative proteins with distinct folds, amino acid compositions, secondary structure, topology and stability. Tests of the model were performed taking 32 proteins from MoDeL that configure the **µMoDeL subset**, which contains representatives of all protein metafolds and covers different classes, topologies and sizes (from N=31 to more than 2500 residues, including both mono and multidomain proteins). Initial training was performed taking 6 proteins of different range size (**1i6f, 1pht, 1agi, 1jli, 1bsn** and **1sur**). The optimized parameters were validated against a test set defined by the remaining 26. To avoid overtraining, the model was further tested in randomly selected proteins of larger size and multidomain (**3adk, 1bud, 1ssx, 1ppo, 1dua, 1qlj, 1pmi**), plus some very large proteins: **1sqc** (619 residues), **1e5t** (710 residues), **1j0m** (747 residues) and **1e9s** (2545 residues). As a final test of the ENMs to represent flexibility in large time scale, results were also compared with those from long MD trajectories (0.1 µsec) in a few illustrative cases: **2gb1, 1ce1, 1cqy** and **1opc**. Benchmarks for comparisons with X-ray and NMR experimental flexibility are described below.

3.2 Molecular Dynamics and Essential Dynamics protocols

Protein structures considered here were titrated, neutralized by ions, hydrated, minimized, heated and equilibrated for at least 0.5 ns (Rueda, Ferrer-Costa, et al., 2007b). Trajectories were collected for at least 10 ns at 300K by using the isothermal-isobaric periodic boundary simulations in explicit water and ions and the Particle Mesh Ewald technique to account for long-rang electrostatic interactions. The quality of MD simulations is dependent on the quality of the force field used, and thus, for each protein 10 ns trajectories were repeated using three all-atoms force-fields (AMBER parm99 (Cornell et al., 1995); CHARMM22 (MacKerell et al., 1998), and OPLS/AA (Jorgensen, Maxwell, & Tirado-Rives, 1996). For computational reasons, trajectories in the 10^2 ns range (or from very large proteins) were collected only with the AMBER force-field. Due to the strong similarity among force fields, when trajectories of the same protein for different force-fields were available they were combined into 30 ns "***force-field independent meta-trajectories***", which provide an averaged view of protein flexibility. The meta-trajectories are compressed through an ED approach (see **Chapter 2**, ***2.1.3***), and the 3N-6 eigenvectors computed only for C-alpha carbons as in coarse-grained ANM. To discard artifacts in these meta-simulations, comparisons were performed also with single force-field trajectories.

3.3 Descriptors for Eigenspaces Flexibility Comparison

The ability of NMA to reproduce MD flexibility can be examined through a variety of metrics that analyze the respective sets of eigenvalues and eigenvectors. Several complementary aspects have been addressed to quantify the degree of similarity between the deformation patterns:

1) Global deformability: The size and complexity of the ''*important*'' deformation space were characterized by different measures, such as i) the **variance**, ii) the **number of modes needed to explain 90% of this structural variance** and iii) the **variance profile,** i.e. how it is distributed along the mode spectra, iv) the **"reduced variance"** defined as the variance explained by the first 5 modes, which for average-sized proteins account for 70-80% of the total ED variance (see *Figure 3. 2*; similar results in (L. Yang et al., 2008) and finally, v) the **strength (force constants) of the softer deformation modes**. Note that in NMA, the hessian eigenvalues are related to mode frequency and thus given in force constants units (kcal/molÅ²). However, ED eigenvalues are structural variances (Å²), which can be converted to stiffness constants assuming each of the 3N-6 modes (N=number of Cα) has an energy k_BT, according to the ***Equipartition Theorem***:

$$K_\nu = \frac{k_B T}{\lambda}$$

(Equation 3. 1 *Mode Stiffness*)

Where ν stands for a given mode, λ_ι stands for the eigenvalue in square distance units, and k_BT is the thermal energy. Accordingly, the variance associated with each NMA mode is given by the inverse of the corresponding force constant multiplied by k_BT.

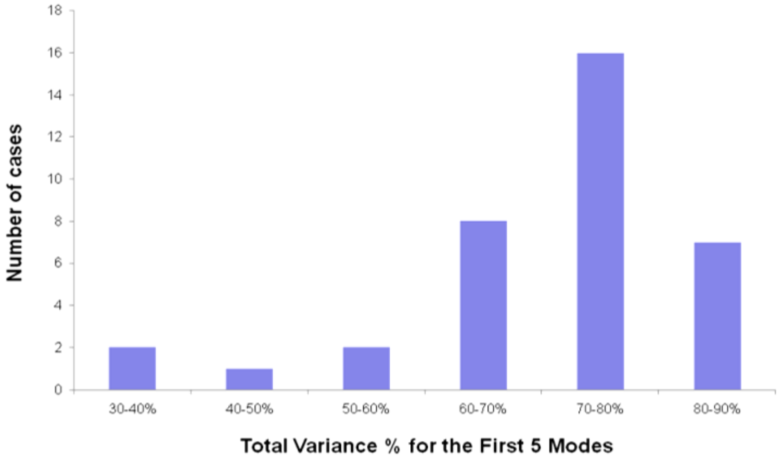


Figure 3. 2 Percentage of total variance captured by the first 5 essential modes in MD trajectories.
For the proteins considered in this study, including the extremely large ones (up to 2545 residues). Note how MD principal components tend to concentrate most of the structural variance (70-80%) associated with collective motions in the first five modes.

2) Deformational space overlap: To evaluate similarity in NMA and ED deformation spaces, the eigenvectors (γ) directions are compared by Hess's metrics (Hess, 2000):

$$\gamma_{AB} = \frac{1}{m} \sum_{\mathrm{i}=1}^{m} \sum_{\mathrm{j}=1}^{m} \left(\left(v_i^A \circ v_j^B \right)^2 \right) \qquad \text{(Equation 3. 2 } \textit{Hess Overlap}\text{)}$$

where *A* and B stands here for the two methods (NMA and ED), the indexes *i* and *j* stand for the orders of eigenvectors and *m* stands for the number of eigenvectors in the "important space". Similarity index in *Eq. 3.2* presents two shortcomings: i) it increases with the size of the important space (for m= *3N - 6* will be equal to one, ii) the index is not sensitive to permutation of eigenvectors (i.e. the same index will be obtained with a perfect 1st ←→1st and 10th ←→10th correspondence than with a 1st ←→10th and 10th ←→1st correspondence). To solve the first limitation we refer Hess's indexes to background models using statistical ***Z-score indexes***:

$$Z_{score} = \frac{\left(\gamma_{AB}(observed) - \gamma_{AB}(random) \right)}{std\left(\gamma_{AB}(random) \right)} \qquad \text{(Equation 3. 3 } \textit{Z-Score}\text{)}$$

Physically-meaningful random models are obtained by diagonalization of a covariance matrix obtained from discrete molecular dynamics simulation, using a limited Hamiltonian consisting only of covalent bonds plus a hard sphere potential at Cα (Emperador et al., 2008). The standard deviation appearing in *Eq. 3.3* is obtained by considering 500 different random models. To evaluate the impact of permutation, we computed dot products between pairs of eigenvectors, determining difference in rank between the eigenvectors showing the largest overlap, and used Perez's index, which weights the similarity of each pair of eigenvectors by their associated Boltzmann's factor (see (Pérez et al., 2005)) for details):

$$\xi_{AB} = \frac{2\sum_{i=1}^{i=z}\sum_{j=1}^{j=z}\left[\left(v_i^A \circ v_j^B \right) \frac{\exp\left\{ -\frac{(\Delta x)^2}{\lambda_i^A} - \frac{(\Delta x)^2}{\lambda_j^B} \right\}}{\sum_{i=1}^{i=z}\exp\left\{ -\frac{(\Delta x)^2}{\lambda_i^A} \right\} \sum_{j=1}^{j=z}\exp\left\{ -\frac{(\Delta x)^2}{\lambda_j^B} \right\}} \right]^2}{\sum_{i=1}^{i=z}\left(\frac{\exp\left\{ -2\frac{(\Delta x)^2}{\lambda_i^A} \right\}}{\left(\sum_{i=1}^{i=z}\exp\left\{ -\frac{(\Delta x)^2}{\lambda_i^A} \right\} \right)^2} \right)^2 + \sum_{j=1}^{j=z}\left(\frac{\exp\left\{ -2\frac{(\Delta x)^2}{\lambda_j^B} \right\}}{\left(\sum_{j=1}^{j=z}\exp\left\{ -\frac{(\Delta x)^2}{\lambda_j^B} \right\} \right)^2} \right)^2} \qquad \text{(Equation 3. 4 } \textit{Weighted Overlap}\text{)}$$

where the common displacement (Δ*x*) is selected as the minimum value, and is negligible the impact outside the important space. Note that Z-score associated to Perez's index is straightforward using *Eq.3.3*.

An additional metrics that helps in determining the similarity between MD and NMA-based eigenvectors is the “spread” index by Hinsen (K Hinsen, 1998):

$$s_{i=}\left(\sum_{j}^{m} j^2\eta_{ij}^2 - \left(\sum_{j}^{m} j\eta_{ij}^2\right)^2\right)^{1/2}$$ (Equation 3. 5 *Mode Spread*)

Where $\eta_{ij} = v_i^A \circ v_j^B$. Note that for two identical sets of modes $\eta_{ij}^2 \neq 0$ only if *i=j* spread becomes equal to 0; higher values indicate the distribution of the eigenvector *i* from space *A* spreads on a larger number of eigenvectors *j* of space *B*.

3) Relative distribution of deformational pattern: The relative distribution of the flexibility along the different residues can be analyzed from different metrics. A powerful one is the “collectivity” index suggested by Brüschweiler (Brüschweiler, 1995), which evaluates the number of residues involved in every essential movement:

$$\kappa_i = \frac{1}{N} exp\left\{-\sum_{n=1}^{N} u_{i,n}^2 log\, u_{i,n}^2\right\}$$ (Equation 3. 6 *Mode Collectivity*)

Where *N* is the total number of residues in the protein, is the mass of each residue and the factor:

$$u_{i,n}^2 = \frac{\upsilon_{i,X}^2 + \upsilon_{i,Y}^2 + \upsilon_{i,Z}^2}{m_n}$$ (Equation 3. 7 *Residue Collectivity*)

are the mass-weighted fluctuations of each residue of mass m_n along mode i.

4) Thermal Fluctuations: B-factors (see *Eq.2.19)* were computed to obtain a measure of residue mobility as in (Atilgan et al., 2001):

$$\langle(\Delta r_i)^2\rangle = \left(\frac{3k_BT}{\xi}\right)[\Gamma^{-1}]_{ii} = \left(\frac{3k_BT}{\xi}\right)\sum_{k}^{m}[\lambda_k^{-1}\upsilon_k\upsilon_k^T]_{ii}$$ (Equation 3. 8 *Residue Fluctuation*)

The correlation between the calculated *B-factors* from ED and NMA was measured by the Pearson and Spearman Correlation Coefficients.

5) Lindemann Coefficients are another useful measure derived from mean fluctuations to evaluate the macroscopic behavior (liquid or solid) of proteins or structural elements (Lindemann, 1910; Rueda, Ferrer-Costa, et al., 2007a; Zhou, Vitkup, & Karplus, 1999):

$$\Delta_L = \frac{\left(\sum_i^N \langle \Delta r_i^{\,2} \rangle / N\right)^{1/2}}{a'} \qquad \text{(Equation 3. 9 Lindemann Index)}$$

Where a’ is the most probable non-bonded near-neighbor distance (taken as 4.5 Å).

To avoid noise introduced by high frequency modes, both B-factors and Lindeman’s Coefficients were computed taking only the first m=50 modes.

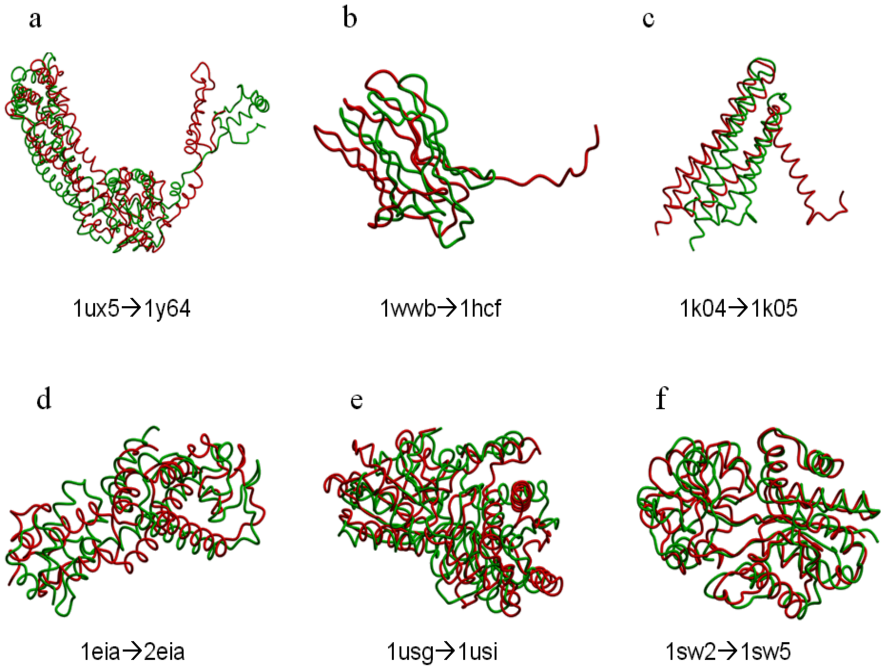


Figure 3. 3. Small benchmark of unbound (red) to bound (green) conformational transitions.

TOP: Large conformational transitions: a) FH2 (formin homology-2 domain) (1ux5→1y64), b) Neurotrophin-binding domain of human TRKB receptor (1wwb→1hfc) , c) Focal adhesion targeting domain of FAK (focal adhesion kinase) (1k04→1k05). BOTTOM: Local conformational changes: d) L-Leucine Binding Protein upon binding with phenylalanine (1usg→1usi), e) Equine Infectious Anemia Virus (EIAV) capsid protein P26 (1eia→2eia), and f) the extracellular ligand-binding protein (ProX) from Archeoglobus fulgidus upon complex with glycine betaine (GB) (1sw2→1sw5). See main text for details.

6) Dot Product against transition vectors and NMR ensembles: The ability of the refined model presented here to trace biologically relevant transitions was tested by computing: *i)* the ***accumulated dot products (Eq.3.2)*** **between the 5 (Overlap (5)) and 10 (Overlap (10)) first eigenvectors** of the structure to the vector driving the transition, and *ii)* the ***rank or distance*** **of the best overlapped eigenvector** (a distance of 0 means is the first one). Systems selected for analysis include three cases displaying large- conformational changes, some of them upon ligand binding: **a)** the **formin homology-2 domain** (1ux5→1y64 transition) (Xu et al., 2004), **b)** the **ligand (neurotrophin) binding domain of human TRKB receptor** (1wwb→1hfc transition) (Ultsch et al., 1999), and **c) focal adhesion targeting domain of focal adhesion kinase** (1k04→1k05) (Arold, Hoellerer, & Noble, 2002); and three other systems showing moderate-local transitions: **d) L-Leucine Binding Protein** (1usg→1usi) (Magnusson, Salopek-Sondi, Luck, & Mowbray, 2004), **e) Equine Infectious Anemia Virus (EIAV) capsid protein P26** (1eia→2eia) (Jin, Jin, Peterson, & Lawson, 1999), and **f) Extracellular binding protein Pro-X from *Archeoglobus fulgidus*** (1sw2→1sw5) (Schiefner, Holtmann, Diederichs, Welte, & Bremer, 2004) (*Figure 3. 3*). To further verify our model, we extended the study to a larger benchmark of 28 open/close conformational changes over 2 Å (comprising 54 structures) from the database **MolMovDB** (Gerstein & Krebs, 1998), and 20 high-quality NMR ensembles for PCA.

3.4 ED-ENM Model Formulation and Parameterization

As we explained before, ENM can be considered a generalization of the *bead-and-strings* Rouse polymer chain model (Rouse, 1953), but contrary to this model where only sequentially adjacent monomers are coupled, ENMs consider that all C^{α}'s within a given threshold are equally connected. Clearly, this is not a realistic approach, since assumes that all interactions within a cut-off are harmonic and identical (irrespective of their chemical nature), and outside are negligible. In order to derive a more physically-sound model we decided to explore alternative approaches. After extensive testing of different connectivity rules, potential functionals and cut-off schemes, we analyzed in detail three models that represent increasing levels of topological complexity and scaling of the constants:

i) A simple **cut-off model** with an uniform constant; this is the most widely used ENM approach (Atilgan et al., 2001; Suhre & Sanejouand, 2004)

$$K_{ij} = C \; if \; r_c > 10\text{Å} \; and$$ (Equation 3. 10 *Cutoff ENM*)

$$K_{ij} = 0 \; otherwise$$

ii) A **non-cut-off model** similar to that developed by Kovacs *et al.* (Kovacs et al., 2004) based on an exponential decay function (Rueda, Chacón, et al., 2007):

$$K_{ij} = C\left(\frac{d_{ij}^{0}}{d_{ij}}\right)^{6}$$ (Equation 3. 11 *Inverse ENM*)

Where d_{ij}^{0}=3.8 Å (C^{α}- C^{α} equilibrium distance) and C=40 kcal/mol.Å^{2}

iii) A **hybrid cut-off model** in which springs for the first M neighbors are weighted according to their sequential distance, while the rest are represented by an optimized exponential decay function (Orellana et al., 2010) (see below)

The proposed hybrid scheme is strongly inspired by the Rouse model, to account for the covalent coupling between chain neighbours. Though there have been attempts to differentiate covalent and non-covalent interactions in ENMs (Kondrashov et al., 2007), they do not consider the covalent peptide backbone coupling and only distinguish *direct* covalent/non-covalent bonds through the scaling of the constants. As we will show, the directions of the main motions are robust to changes in the values given to the springs. Therefore, approaches based on selecting the constants according to the residue identity have not been as successful as expected. Our model defines optimal values for the constants that give ED-closest variances, but also translates the covalent/non-covalent differentiation as topological information.

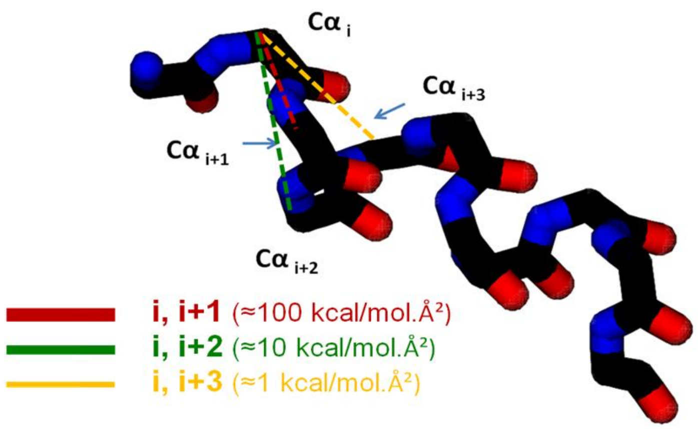


Figure 3. 4. The ED-ENM model as a nearest-neighbors model.

The ED-ENM is a nearest-neighbours based model, maintaining the secondary structure stereochemistry, where the three first order constants acquire values close to a 100:10:1 ratio.

To properly scale the bonded and non-bonded interactions, we connect the first three neighbors independently of the Cartesian distance between the C-alpha of the residues, with a sequential-distance dependent function (*Figure 3. 4*). Beyond the first three relevant neighbours, the elastic potential is based on distances between the Ca atoms that define the covalent skeleton of the peptide chain.

In order to obtain spring weights for the first 1-3 sequential neighbours in an unbiased way we computed the dependence of topological-linked residue-residue "apparent" stiffness constants from MD (see *Eq. 3.12* and *Figure 3. 5*) and fitted them to a inverse exponential function of the sequential distance (*Eq. 3.13*) using a non-linear regression routine for a small set of proteins:

$$K_{ij}^{app} = \frac{k_B T}{\langle (d_{ij} - d_{ij}^0)^2 \rangle}$$

(Equation 3. 12 *Apparent Force Constant*)

Where k_BT is the thermal energy, and R_{ij} and R_{ij}^{0} are the instantaneous and equilibrium distances between any residue pair *i, j*. From the MD data, the dependence between sequential distance and interaction strength can be adjusted to a function of the form:

$$K_{ij}^{app}(S_{ij}) = \frac{C_{seq}^{app}}{S_{ij}^{n_{seq}}}$$

(Equation 3. 13 *Apparent sequential Force Constant*) where S_{ij} stands for a dimensionless distance in sequence between residues *i* and *j*, and C_{seq}^{app} has units of kcal.mol/Å^2. The optimum exponent determining the shape of the variation is used in the rest of the study, while the constant (C_{seq}) is further refined to reproduce "real" instead of "apparent" force-constants. A similar strategy was also used to derive the distance-dependence of weighting constants for non-sequential interactions, obtaining a relation for the force-constants with the Cartesian distance:

$$K_{ij}^{app}\left(d_{ij}\right) = \frac{C_{cart}^{app}}{d_{ij}^{n_{cart}}}$$ (Equation 3. 14 *Apparent Cartesian Force Constant*)

where d_{ij} is the distance in Å between residues *i* and *j* in the native conformation, and C_{cart}^{app} has units of kcal.mol.Å$^{\text{ncart-2}}$. An additional size-dependent cut-off, r_C, was introduced to further annihilate artefactual distant interactions (see below). Then, we define network topology by a hybrid matrix equal to the sum of a Rouse Chain topology matrix for the first *M* neighbours plus a Kirchhoff matrix ***Γ*** for distant interactions, rendering a mixed connectivity matrix that both combines sequential and distant information. Thus, given a pair of residues *i*, and *j* with sequential distance S_{ij} > 0 and Cartesian distance d_{ij}, the ij-th element of the inter-residue contact matrix is:

$$\Gamma_{ij}\begin{cases} S_{ij} \le M, \quad \Gamma_{ij} = 1 \\ S_{ij} > M, \begin{cases} \Gamma_{ij} = 1 \quad if \quad d_{ij} \le r_c \\ \Gamma_{ij} = 0 \quad otherwise \end{cases} \end{cases}$$ (Equation 3. 15 *ED-ENM Topology matrix*)

Where Γ_{ij} is a Kirchhoff matrix with non-zero hepta-diagonal entries defining neighbor sequential contacts and the usual cutoff background for distant interactions. The submatrices for diagonal elements (S_{ij} =0) are defined as in *Eq.2.18*:

$$\Gamma_{ii} = -\sum_{k,k\neq i}^{N} \Gamma_{ik}$$ (Equation 3. 16 *ED-ENM Diagonal Elements*)

Thus, the force-constant γ associated to any residue pair *i, j* is not uniform but dependent on the Cartesian and the sequential distance between them. In terms of the stiffness matrix, $K_{ij}=\gamma\Gamma_{ij}$:

$$K_{ij}\begin{cases} S_{ij} \le M, \quad K_{ij} = C^{seq} \big/ S_{ij}^{n_{seq}} \\ S_{ij} > M, \begin{cases} if \quad d_{ij} \le r_c \quad then \quad K_{ij} = \left(C^{cart} \big/ d_{ij} \right)^{n_{cart}} \\ K_{ij} = 0 \quad otherwise \end{cases} \end{cases}$$ (Equation 3. 17 *ED-ENM Stiffness matrix*)

Where the exponents n_{seq} and n_{cart} and the magnitude of the effective force constants, C_{seq} (in kcal.mol/Å^2 units) and C_{cart} ((kcal.mol/Å)$^{1/ncart}$ units) are obtained by fitting the apparent force-constant to ED variance profiles for a small set of training proteins. The cut-off radius, r_c, was found to be approximately size-dependent in a range of 8-14 Å. An $M = 3$ limit for sequential interactions was determined from MD apparent constants and tested in networks where these contacts are switched on and off.

3.5 Optimization of the Method against the MoDeL Training Set

As described above, we used MD results in a few training proteins to refine the key elements of the model, testing it later in a larger set. The elements to analyze in the training part of the study were: i) *the functional for the dependence of the force-constant with the distance*, ii) *the relevance of sequential and spatial relationships*, iii) *the optimal cut-off for distant interactions*, and iv) *the magnitude of the constants* (C^{seq} and C^{cart}) to be used in the calculation of the effective force-constants (see *Eq. 3.13-3.14*). Since a multi-parametric fitting of all these elements to MD might yield to an over-trained method without physical sense, we decided to follow a conservative stepwise optimization strategy to guarantee the generality of the method.

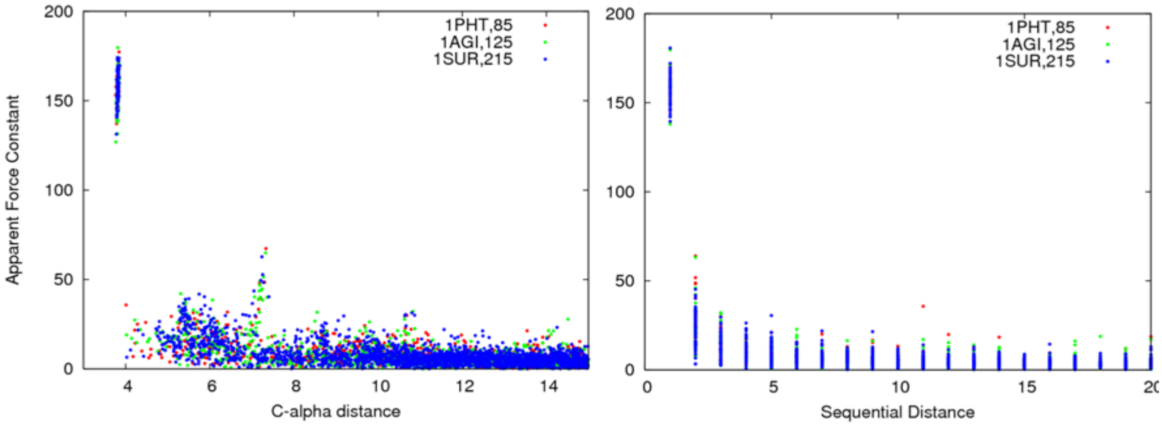


Figure 3. 5. Dependence of the apparent residue-residue force constant with distance.
In Cartesian and sequence space as determined for MD simulations of the training proteins. Results shown correspond to three proteins of different size from the training set: 1PHT (red), 1AGI (green) and 1SUR (blue). Note how first, second and third order neighbor contacts are well defined against the background interactions.

3.6.1 Pseudo-Harmonic Constants Dependence on Cartesian and Sequential distance

The first step in our optimization procedure was to analyze the MD simulations (see *Eq. 3.11*) to determine reasonable functions for the distance-dependence of apparent inter-residue force-constants. As can be seen (*Figure 3. 5*, left), in the limit of uncoupled oscillators the apparent force-constants decay exponentially with C_α-C_α distance, with evident deviations at distances corresponding to *i→i+1* residue interactions (around 3.8 Å) and to a lower extent at distances corresponding to *i→i+2* and *i→i+3* sequential interactions (around 6 and 10 Å). The differential nature of *i→i+1* interactions (and to a lower extent *i→i+2* and *i→i+3*) becomes evident when the dependence of the apparent force-constant with the sequential distance is represented (*Figure 3. 5* right). Fitting of force-constants to sequence distance for close chain neighbours reveals an approximate order two exponential relationship (n_{seq}=2 in *Eq. 3.13*). However, the pre-exponential factor (C^{seq}) in *Eq. 3.13* cannot be taken directly from MD apparent force-profiles, but must be re-fitted to avoid over-restriction of the protein movement, resulting from indirect interactions if effective force-constants were set equal to apparent ones (see section 3.6.3).

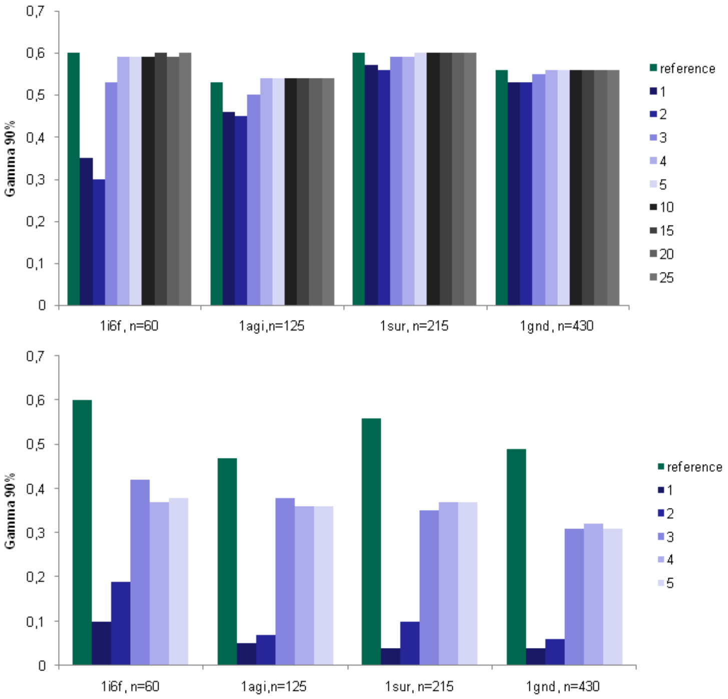


Figure 3. 6. Coupling and uncoupling of interacting residues in protein networks. TOP: similarity index (γ 90% variance) between ENM and ED eigenvectors, when in the ENM sequential constants i, i+j with j=[1,25] (right labels) were deleted keeping a background continuous network connecting all pairs within a cutoff with a distance-dependent spring. BOTTOM: similarity index (γ 90% variance) with ED eigenvectors in a minimal network connecting only close neighbors i, i+j with j=[1,5] with sequentially-weighted constants. The reference is a fully connected network as in the Kovacs formalism.

3.6.2 Flexibility Patterns for changing topologies and connectivity rules

The rules to assign connectivity status to a given pair of nodes are usually based on the Cartesian distance. To further prove that the weight of the first *j*=3 neighbors is critical and to better determine the influence of the sequential distance (S_{ij}), we built test networks in which sequential couplings are switched on and off. Deletion of distant sequential interactions in continuous networks, such as the Kovacs model, does not affect overlap with ED, whereas contacts with close neighbors (specially i, i+2 and i+3) appear to be critical for the main modes directions (see *Figure 3. 6*, top). This leads us to question which accuracy of prediction could reach a true Rouse-chain model. We built a minimalist ENM, where only sequential-level i→i+1, ..i+3 interactions are included and all distant contacts are ignored. Not surprising, just coupling the first three neighbors the minimal model can reproduce the deformation patterns obtained more complex models, reaching 50% or more of the maximum overlap with ED obtained with a fully connected reference network (*Figure 3. 6*, bottom).

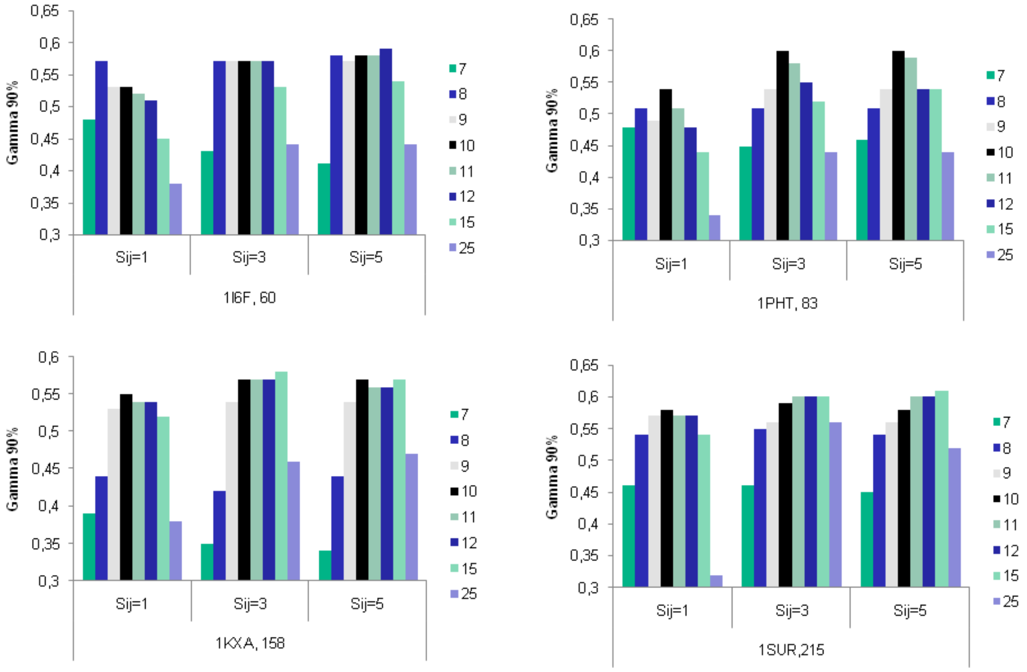


Figure 3. 7. Similarity index in minimalist nearest-neighbors networks.
(Gamma 90% variance) between ENM and ED eigenvectors when ENM is computed considering simple 1/0 networks where only close sequential constants i, i+j with j=[1, 5] are weighted over the background. Cutoff range is also explored from 7-25 Å (right labels).

The order-two exponential function weights the relative strength of the interactions between these three nearest-neighbours as **:100:10:1**, a proportion crucial for mode directionality. The possibility to use other definitions of the chained residues (including more or less terms) was further analyzed in simpler networks where an increasing range of sequential contacts was weighted over a cutoff background, confirming again the i, i+3 limit for main-chain interactions. For example, the 1-neighbour sequence list proposed as a minimum requirement for mechanical stability by Jeong (Jeong et al., 2006), and topologically equivalent to a constants scaling of **100:1:1**, gives suboptimal results. As shown in *Figure 3. 7*, i→i+2 and i→i+3 contacts must be clearly weighted over the background, and a cutoff larger than 8 Å has to be considered for non-bonded contacts. By introducing stronger force constants between close chain-neighbours, the backbone motions around torsional angles are restrained to near-equilibrium stereochemically allowed conformations. On the contrary, extension of the chained residues to i→i+5 interactions did not yield any improvement, as could be already anticipated from *Figure 3. 5*. All these findings strongly suggest that interactions between close chain neighbours, defining the local covalent topology, are the key to define soft modes directions and that protein behave in practice as networks of reduced connectivity regarding its dynamics.

3.6.3 Topology of interactions and threshold distance

Analysis of *Figure 3. 5*. and inspection of the essential dynamics of training trajectories reveals that there is a threshold distance from which the apparent restriction in the movement of two pairs of residues is very small and can be fully explained by indirect interactions (see *Figure 3. 8*), without the need to include direct interactions. This recommends the use of a cut-off to eliminate artefactual restrictions to movement due to unrealistic distant interactions and to reduce computational cost in large proteins.

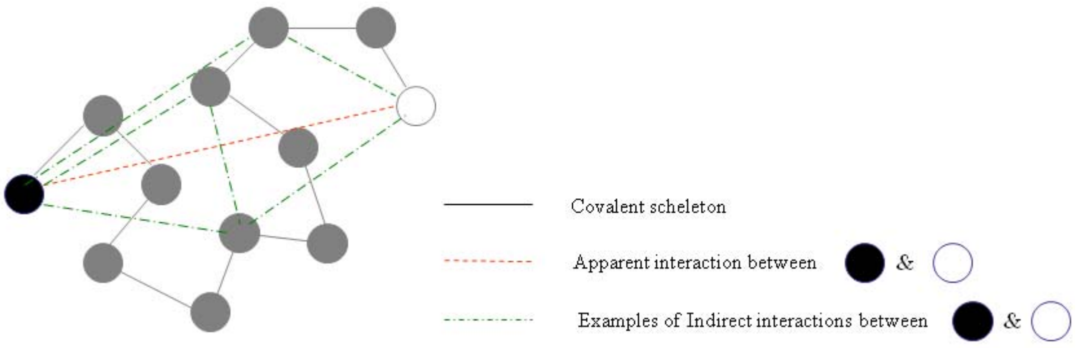

Covalent scheleton
Apparent interaction between ● & ○
Examples of Indirect interactions between ● & ○

Figure 3. 8. Schematic representation of direct and indirect interactions between residues.
The difference between apparent and real interactions between two particles is illustrated.

Optimum cut-offs were determined by analyzing the overlap (*Eq. 3.2*) between ENM-based NMA and MD essential deformation modes and the relative variance profile, which are independent of the magnitude of force constants. For most average-density globular proteins the optimum cut-off is roughly proportional to the length of the protein as given by the number of residues, N (see *Figure 3. 9*):

$$r_c = int(3\,log\,N - 2.8), \qquad if\ N > 50\ and$$

$$r_c = 8\ otherwise$$

(Equation 3. 18 *Cutoff radius*)

A relation used in the following to determine the cut-off radii for our model in the remaining calculations. Exceptional deviations from this general rule are found for proteins with extreme packing densities or ellipsoidal shapes. However, it must be noticed that our model is not cutoff-dependent, and the precise election of the threshold distance within the size-cutoff range, does not affect the results as the main topological information is encoded by the nearest-neighbours sequential connectivity.

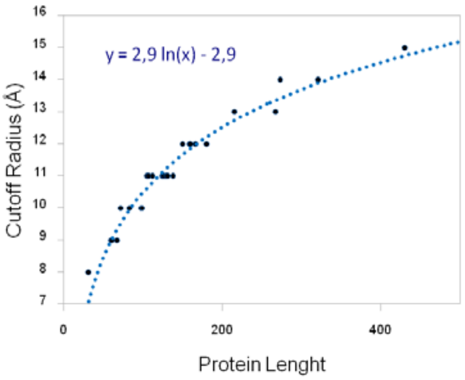


Figure 3. 9. Optimum cutoffs (in Å) depending on the length (in number of residues) of the protein. We define the optimum cutoff as the distance r_{max} that gives best overlap with ED flexibility descriptors for the range r_{max}+/-1 Angstrom. The boundaries for optimal cutoff assignation are between 8-20 Angstroms.

3.6.4 Springs magnitude: Robustness of mode directionality

When sequential interactions are removed from *Figure 3. 5* the apparent force-constants accounting for long-range non-covalent interactions are found to decay with the cartesian distance following an order 6 exponential (n=6 in *Eq. 3.13*), which matches that suggested by Kovacs (Kovacs et al., 2004). Such dependence was directly incorporated in the method, whereas the pre-exponential factor C^{cart} is fitted to improve the overlap with variance plots from MD trajectories. Once the ideal functional forms were determined, we fitted the force-constants by comparison with ED-MD estimates of: i) *total variance*, ii) *variance profile*, and iii) *variance of the first five modes*. We explored systematically values for the effective sequential C^{seq} (in the range 40-200 kcal/molÅ^2) and Cartesian C^{cart} (in the range 2-12 kcal/molÅ^2) constants finding optimal agreement for C^{seq} =60 kcal/molÅ^2 and C^{cart}= 6 (kcal/molÅ) $^{1/6}$. It is worth noting that these results are quite robust to changes of ±10 kcal/molÅ^2 in C^{seq} and ±1 (kcal/molÅ) $^{1/6}$ in C^{cart} (see *Figure 3. 10*). The direction of the soft modes, as reflected by Hess metrics, is not sensible to the magnitude of the force constants.

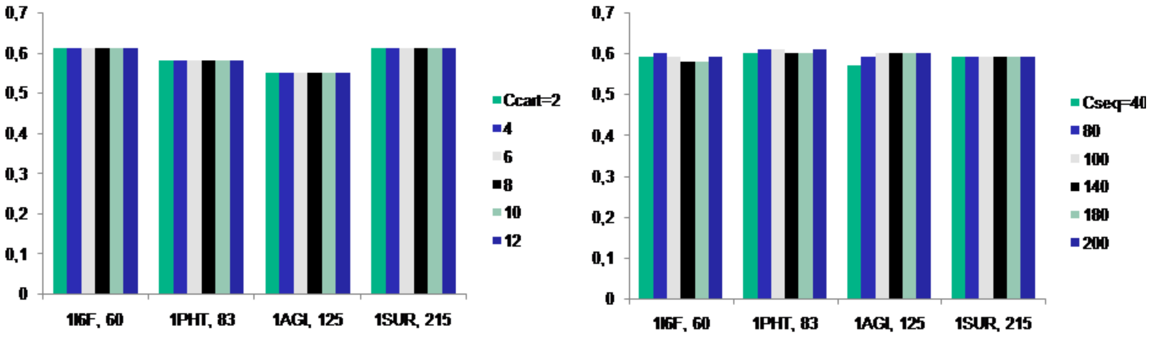


Figure 3. 10. Examples of the robustness of ED-ENM modes directionality to changes in C_{seq} and C_{cart}. TOP: changes in similarity index (γ 90% variance) with C_{seq} in a 1/0 network where only first 3 neighbours are weighted according to equation 6 (M=3, n_{seq}=2, size-dependent cutoff). BOTTOM: changes in similarity index (90% variance) with C_{cart} in a 1/0 network where contacts are weighted according to continuous function in (n$_{cart}$=6, size-dependent cutoff). The change of C_{seq} and C_{cart} introduces larger changes in total and mode variances.

3.6 Validation of the method.

3.7.1 Comparison of the Mixed Algorithm with standard ANM Methods

As described before, we analyzed the behavior of the new model by comparing our ED-ENM with ED analysis of MD meta-trajectories (Rueda, Chacón, et al., 2007). We also included as reference comparisons with the most widely used linear cutoff-method and with the six-power exponential function developed by Kovacs et al. In all cases calculations were performed using the MD-averaged structure as reference to allow us a direct comparison between coarse-grained NMA methods and atomistic MD simulations. All ENMs considered here show a reasonably good ability to reproduce the flexibility pattern determined by MD simulations. Average results on the test set displayed in *Figure 3. 11* illustrate that any of the ENMs is actually displaying a reasonable ability to reproduce the MD samplings. We computed the dot product or similarity index (*Eq. 3.2*) and their Z-score (*Eq. 3.3*), to evaluate the similarity of the directions in the eigenvectors from ED and NMA. Similarity indexes for 90% variance are in the range 0.5-0.6 (0.6-0.7 if index is computed considering always 50 eigenvectors), with highly significant associated Z-scores (around 100). The large Z score values indicate that all similarity measures are far from random noise. These similarity indexes are not far from those obtained by comparing MD trajectories obtained with different force-fields (**0.7-0.8**) ((Emperador et al., 2008); see *Table 3. 2*).

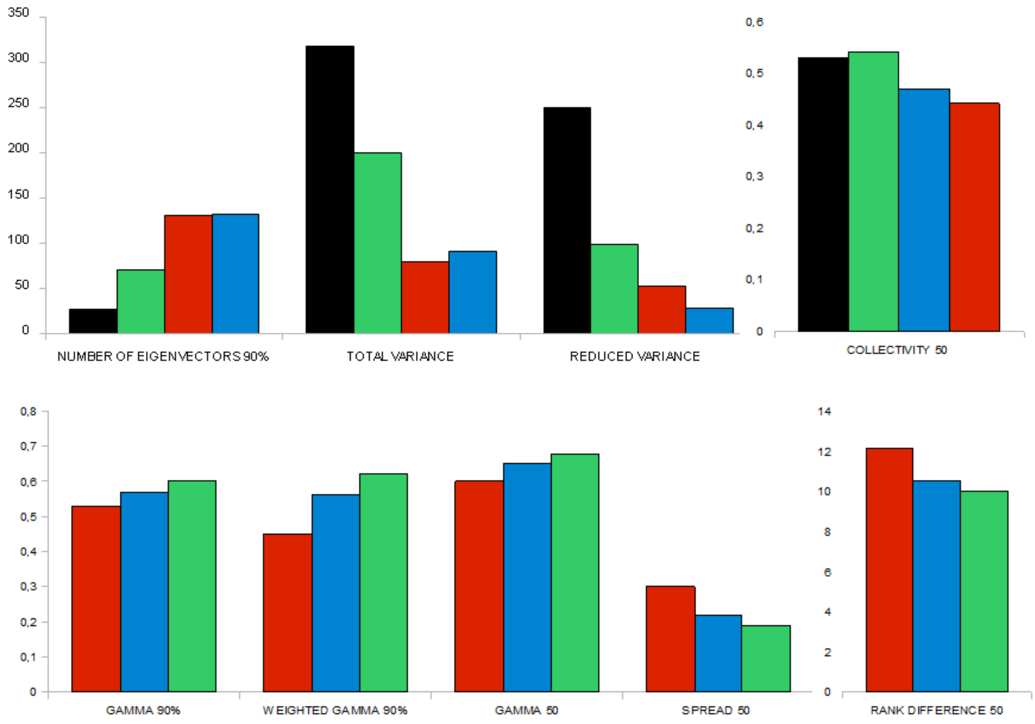


Figure 3. 11 Different metrics for the comparison between MD and ENM-NMA.
Color code in this and the next figures: *Black:* reference MD simulations *Green:* ED-ENM model, *Red:* standard cut-off model and *Blue:* Kovac's formalism.

However, despite the general good agreement it is clear that not all the ENM presented here show the same performance: in general, the simple cut-off method fails in difficult cases where stiffer interactions may block the rearrangements required for a large-scale movement, whereas our model yields the better results in most proteins, with near-top similarities in all cases (see *Table 3. 1*, *Figure 3. 11*). The ED-ENM model leads to a moderate, but significant increase in the average similarity index respect to the cutoff/inverse models (from **0.54/0.57** to **0.60** in γ (90%) and **0.61/0.65** to **0.68** for γ (50)), which reflects that the type of movements sampled in MD are closer to ED-ENM calculations. Interestingly, the improvement is focalized in the most prevalent eigenvectors, as shown by the increase in the weighted similarity index for the γ90 (from **0.45/0.56** to **0.62**) (see *Table 3. 1*). When exploring the entire similarity index space, it becomes clear that the mixed model improvement in similarity indices is due preferently to the increase in the Dot Product between the first ten eigenvectors, which are at the same time the most representative of the collective motions. This improvement is more noticeable for medium-small proteins, which gave the worst results with the standard models. The close correspondence between MD and ED-ENM deformation movements becomes also clear in the corresponding "*spread*" of each NMA eigenvector in ED modes, which is smaller (*Figure 3. 11)* especially considering the first lowest frequency eigenvectors (*Figure 3. 12*).

Table 3. 1 Comparative Measurements of Flexibility Patterns Obtained with NMA and ED.

PDB code (CATH)	Total Variance* Eigenvectors 90% Variance*	Similarity $(\gamma_{(10)})^{\#}$	Similarity $(\gamma_{(90\%)})^{\&\#}$	Z-score (90% var) $^{\#}$	Similarity $(\gamma_{(50)})^{\#}$	Z-score (50 eig) $^{\#}$	Pearson Coeffic.$^{\#\$}$
1OPC	201/67/56/140	0.46/0.49/0.48	0.56/0.59/0.61	26/29/31	0.63/0.68/0.70	93/104/109	0.50 / 0.59 / 0.65
99 (α)	19/46/96/44		0.49/0.60/0.60				0.33 / 0.25 / 0.39
1CSP	86/45/46/73	0.51/0.54/0.61	0.62/0.64/0.68	37/39/44	0.64/0.70/0.72	64/75/79	0.46 / 0.55 / 0.71
67 (β)	20/30/61/38		0.61/0.68/0.72				0.49 / 0.54 / 0.62
1SDF	460/76/92/556	0.48/0.53/0.53	0.43/0.43/0.49	23/23/28	0.66/0.63/0.67	48/43/50	0.76 / 0.77 / 0.79
67 (α+β)	7/15/38/9		0.16/0.22/0.52				-
1OOI	131/38/53/103	0.28/0.36/0.40	0.59/0.66/0.68	20/34/38	0.63/0.71/0.72	127/149/151	0.40 / 0.61 / 0.60
124 (α)	37/131/133/74		0.47/0.21/0.69				0.23 / 0.46 / 0.65
1BFG	85/27/52/75	0.44/0.49/0.51	0.62/0.67/0.71	37/50/61	0.62/0.66/0.70	145/158/170	0.39 / 0.58 / 0.59
126 (β)	54/166/143/94		0.66/0.73/0.74				0.30 / 0.30 / 0.50
1CHN	359/138/71/160	0.46/0.47/0.49	0.48/0.52/0.53	19/23/24	0.61/0.66/0.68	131/146/151	0.54 / 0.68 / 0.74
126 (α+β)	15/29/118/62		0.38/0.52/0.55				0.35 / 0.62 / 0.53
1IL6	840/43/105/252	0.50/0.50/0.49	0.49/0.50/0.50	27/28/28	0.60/0.66/0.66	95/109/109	0.68 / 0.81 / 0.83
166 (α)	9/164/139/77		0.09/0.28/0.43				-
1CZT	197/42/112/146	0.42/0.49/0.49	0.58/0.65/0.69	42/54/61	0.60/0.65/0.69	111/124/134	0.51 / 0.56/ 0.72
158 (β)	38/140/140/97		0.54/0.70/0.72				0.66 / 0.67 / 0.77
1GND	1022/83/248/484	0.45/0.51/0.51	0.53/0.56/0.58	23/25/27	0.56/0.61/0.62	330/363/370	0.75 / 0.77 / 0.72
430 (α+β)	30/521/409/214		0.27/0.32/0.65				0.48 / 0.57 / 0.53
1BR5	185/47/150/274	0.40/0.44/0.45	0.62/0.68/0.68	41/59/59	0.58/0.64/0.64	200/225/225	0.65 / 0.71 / 0.73
267 (α)	85/353/261/146		0.56/0.73/0.72				-
2PIA	255/69/210/364	0.54/0.59/0.60	0.60/0.65/0.66	33/43/46	0.57/0.62/0.62	170/189/189	0.55 / 0.60 / 0.62
321 (β)	96/366/305/162		0.56/0.63/0.71				0.49 / 0.52 / 0.50
2HVM	376/32/112/183	0.41/0.45/0.45	0.55/0.61/0.60	33/43/42	0.56/0.62/0.61	177/200/196	0.68 / 0.84 / 0.81
273 (α+β)	44/449/307/184		0.27/0.55/0.61				-

* Values in the cells correspond always to: MD/cut-off NMA/Kovac's/ED-ENM method. & Values in the first line of the cells correspond to the standard Hess's metrics (*Eq. 3.2*) and values in the second line to Perez's index (*Eq. 3.4*). In every line results displayed correspond to cut-off NMA/Kovac's/ED-ENM method. # Values in the cells correspond to cut-off NMA/Kovac's/ED-ENM method. $ Values in the first line of the cells correspond to correlations against ED atomic fluctuations, and values in the second line to correlations against experimental B-Factors. In every line results displayed correspond to cut-off NMA/Kovac's/ED-ENM method.

Analysis of total variances and variance profiles reveals some of the most serious shortcomings of the standard ENM-based NMA models, which largely underestimate the total variance with respect to MD simulations (by a factor of 3-4 when standard cut-off and Kovac's methods are used). Note that this deviation cannot be fully explained by the fact that we are using a MD meta-trajectory as reference, which might include some variance differences associated to each force-field (see *Table 3. 2*). Very interestingly, the deviation in variance between ENM-NMA and MD simulations is not uniform for the entire deformation space (which will make easy the correction by scaling down residue-residue force-constants), but it is larger for the first essential movements, as shown in the fact that the "*reduced*" variance (i.e. that of the first 5 eigenvectors) in standard ENM is 5-8 times smaller than that computed by atomistic MD simulations. The MD deformation space that is bigger (in terms of variance) is also less complex (i.e. less eigenvectors are required to explain a given variance threshold) than the deformation space obtained from standard ENMs (see *Table 3. 1* and *Figure*

3.13). The reason for this behavior is clear from the analysis of the variance profiles and of the force-constants (K_v in *Eq. 3.1*) associated to essential deformations, which illustrates the very different way in which standard NMA and MD simulations distribute variance along the intrinsic deformation patterns (see *Figure 3.13*). Thus, while MD defines a small number of movements which concentrate most of the variance, the deformability is distributed along a larger number of eigenvectors in standard ENM. In summary, not only the total variance is different, but MD and standard ENM differ also in how this variance is partitioned between modes, and this is something that cannot be modified by modifying a universal spring constant.

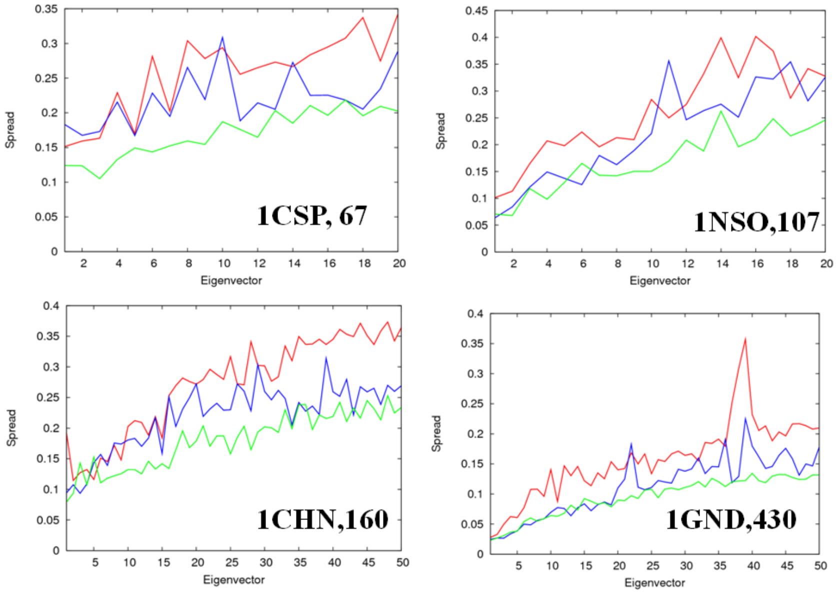


Figure 3. 12. Spread of the Eigenvectors in the ED eigenspace for randomly selected proteins.
The standard ENM approaches split the MD important space in a greater number of eigenvectors. Color code is as in Figure 3.11.

The hybrid ED-ENM presented here is still a NMA-based approach and accordingly cannot be expected to fully capture non-harmonic MD movements which are the main responsible for the softer deformation modes in atomistic trajectories and accordingly, for the large MD variance. However, the method yields to a clear improvement in the total variance and, more important, in the balance of deformation movements (*Table 3. 1* and *Figure 3. 11-Figure 3.13).* Thus, the deviation from MD in total variance is only by a factor of 1.6 (2.5 in “reduced” variance) and the complexity of the deformation space decreases by a factor of 5 with respect to the other ENM. The variance and force-constant (K_v in *Eq. 3.1*) profiles are also in much better agreement with those derived from MD simulations. It is worth to note that the improvement in the representation of the MD essential deformation space obtained by using the ED-ENM model is constant for all the range of proteins considered and for all structural families,

Table 3. 2 Comparative Measures of Flexibility Patterns Obtained with different MD force-fields and our mixed ENM model for some representative proteins.

PDB	*Force-Field*	*Total Variance**	*Eigenvec. 90% Var*	*ACO[1]/AMBER/CHARMM/OPLS Similarity (90% var)[&]*	*ACO/AMBER/CHARMM/OPLS Similarity (50 eig)[#]*
1OPC	ED-ENM	137(62,45%)	43	0.58/0.59/0.68/0.62	0.68/0.67/0.69/0.67
	ACO	212(149,70%)	22	1/0.74/0.75/0.70	1/0.82/0.82/0.80
	amber	132(85,64%)	29	0.69/1/0.65/0.57	0.82/1/0.73/0.70
	charmm	99(58,59%)	35	0.69/0.60/1/0.64	0.82/0.73/1/0.74
	opls	112(74,66%)	30	0.65/0.56/0.68/1	0.80/0.70/0.74/1
1OOI	ED-ENM	102(30,30%)	73	0.64/0.68/0.67/0.66	0.70/0.69/0.71/0.69
	ACO	152(94,62%)	35	1/0.80/0.80/0.81	1/0.81/0.84/0.83
	amber	101(61,61%)	44	0.77/1/0.68/0.70	0.81/1/0.74/0.73
	charmm	130(82,63%)	35	0.80/0.72/1/0.71	0.84/0.74/1/0.75
	opls	107(67,63%)	39	0.78/0.72/0.69/1	0.83/0.73/0.75/1
1CZT	ED-ENM	146(44,30%)	97	0.66/0.65/0.66/0.64	0.66/0.65/0.66/0.64
	ACO	236(152,64%)	37	1/0.77/0.78/0.78	1/0.77/0.78/0.78
	amber	139(79,57%)	48	0.74/1/0.67/0.67	0.77/1/0.67/0.67
	charmm	150(86,57%)	51	0.75/0.66/1/0.67	0.78/0.66/1/0.67
	opls	142(78,54%)	49	0.77/0.67/0.68/1	0.78/0.67/0.68/1
2HVM	ED-ENM	181(38,21%)	184	0.58/0.67/0.63/0.62	0.57/0.59/0.55/0.57
	ACO	384(268,70%)	54	1/0.80/0.79/0.77	1/0.70/0.71/0.72
	amber	148(57,39%)	124	0.71/1/0.66/0.65	0.70/1/0.55/0.57
	charmm	211(117,55%)	93	0.72/0.70/1/0.65	0.71/0.55/1/0.57
	opls	234(137,58%)	79	0.73/0.72/0.68/1	0.72/0.57/0.57/1

[1] *ACO=AMBER-CHARMM-OPLS metatrajectories*

* Values in the cells correspond to: Total Variance (Reduced Variance, %)

[&] Values obtained considering standard Hess metrics (*Eq.3.2*) with a number of evec representing 90% of the Variance for the corresponding reference Force-field (top labels in columns 3-4) as shown by selected examples in *Table 3. 1*. The amount of residues involved in essential deformation movements is quite similar in ENM and MD, as noted in collectivity measures (see collectivity for the first 50 modes, *Figure 3. 11*), but even small there is a uniform tendency of standard methods to less collective movements than those detected in MD. Such a tendency is corrected in our ED-ENM method, suggesting that residue mobility is more realistic in hybrid calculations than in the other two standard ENM approaches. This suggests that collective motions are more cooperative in ED-ENM, possibly due to the strongest nearest-neighbors coupling.

Table 3. 3 Statistical Correlation Coefficients of the B-factor distributions obtained from NMA calculations and reference values (MD or X-ray)

REFERENCE DATA	MD	CUT-OFF	KOVACS	ED-ENM
X-ray B-factors	0,51±0,11	0,46±0,12	0,55±0,12	0,56±0,12
MD simulations	-	0,52±0,14	0,63±0,16	0,64±0,14
MD in the X-ray subset (21)		0,53±0,14	0,63±0,11	0,65±0,10
MD in the NMR subset (10)		0,48±0,12	0,64±0,23	0,58±0,15

* NMA and MD (ED) values were computed always from the mean-square fluctuations obtained when activating the first 50 eigenvectors.

Projection of the collective modes on individual residues allowed estimating residue fluctuations in solution (*Eq.3.8*). As previously reported all ENMs reproduce (see *Table 3. 3*) the MD atomic fluctuations reasonably well, with Pearson's correlation factors in the range 0.5-0.6 (typically 0.7-0.8 Spearman's coefficients). But, when individual fluctuations distributions are compared (see *Figure 3. 14)* the shortcomings of standard ENMs become evident in a flattening of the B-factor profiles, resulting from inability of ENM to capture local but large non-harmonic deformations. It is also worth to note that, even our interest was not on the description of flexibility in the crystal but in solution, the ED-ENM approach yields also a sizeable in the X-ray B-factor profiles. We also found that it is possible to raise the correlations for B-Factors by increasing the distance threshold (*unpublished data*), but as a result the structure becomes stiffened and accuracy decreases in other global flexibility measurements, such as similarity index, variance profiles, or the ability to trace large conformational changes as we will discuss below.

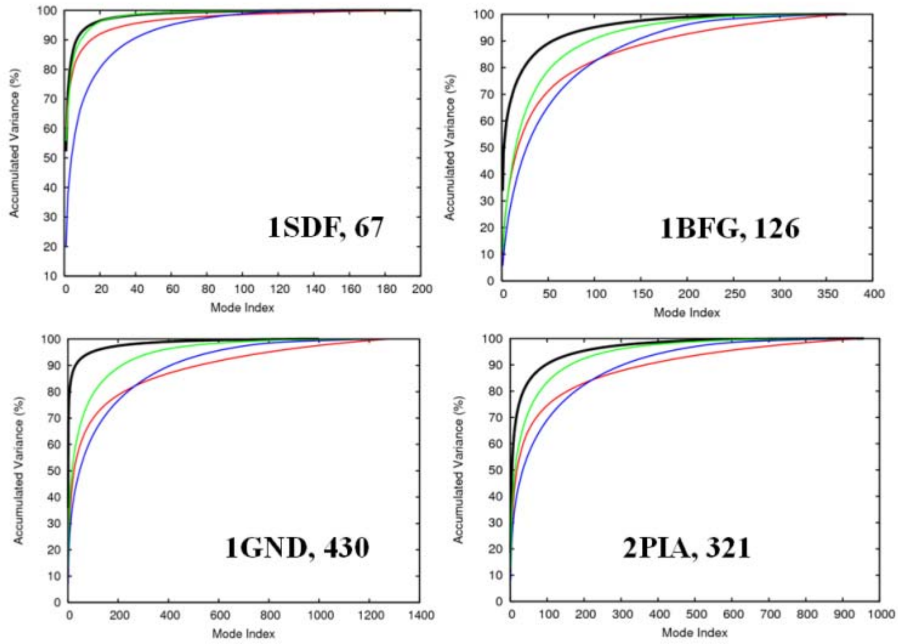
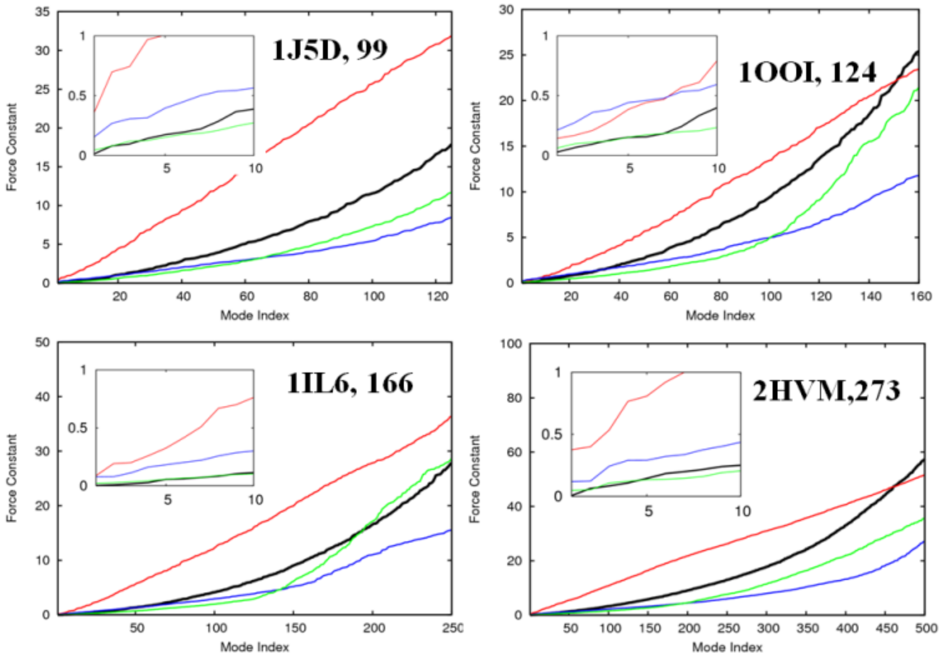


Figure 3. 13 Percentage of Accumulated Variance and Force Constants for random proteins.

(A) ABOVE: Cumulative variance with respect to the number of eigenvectors and (B) BELOW: Strengths of the essential deformation modes (K_v, *Eq. 3.1*) computed by the different methods for a selected number of typical proteins (insert correspond to a zoom of first eigenvalues. Illustrative proteins of different sizes and secondary structure compositions are displayed (the name and number of residues is shown). Color code is as in Figure 3.11.



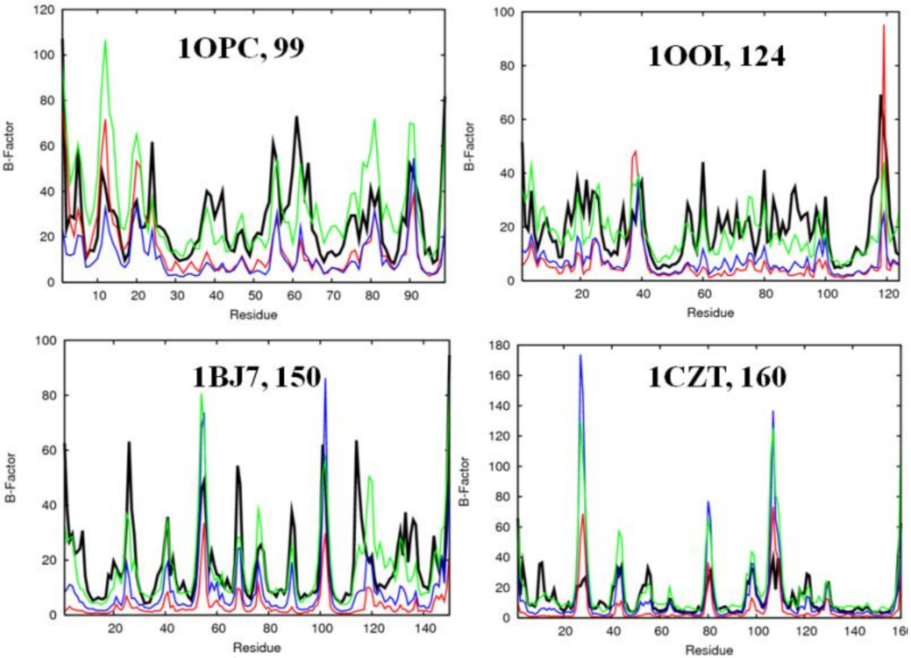


Figure 3. 14. B-factor profiles (Å^2) computed by the different methods for a selected number of typical proteins.

Illustrative proteins of different sizes and secondary structure compositions are displayed (the name and number of residues is shown). Values obtained considering in all cases movements along the first 50 eigenvectors. Note the flattening of standard ENM profiles compared with MD. Color code is as in Figure 3.11.

A simple post-processing of B-factors allows the derivation of Lindemann's index (see Methods) a key descriptor to analyze the macroscopic nature of proteins. The results in *Table 3. 4* illustrate the superiority of the ED-ENM method with respect to the standard ENMs to reproduce the absolute MD-derived Lindemann's index. It is also worth noting that the ED-ENM method reproduces nicely the inside/outside (solid/liquid) asymmetry of proteins and the different macroscopic behavior of the three kind of secondary structures considered here. That indicates that our ED-refined ENM method provides a reasonable and improved description of the solid/liquid macroscopic nature of the different protein regions.

Table 3. 4 Lindemann's coefficients* for different secondary structure elements and residue positions in MD and NMA calculations.

	MD	Cut-off	Kovacs	ED-ENM
All residues	0.24	0.15	0.15	0.25
Buried	0.19	0.09	0.11	0.18
Exposed	0.26	0.17	0.17	0.28
α-helices	0.21	0.10	0.12	0.22
β-sheets	0.17	0.11	0.11	0.19
Turns	0.28	0,20	0.19	0.28

***** Averaged values for MICROMODEL database.

All the results reported to this point suggest that the ED-ENM is able to provide reasonable approximations to MD, improving significantly the results obtained by current ENMs without increasing formal or computational complexity. There are however, two reasons for concern regarding the behavior and transferability of the model in the biologically relevant time and length scales, namely: i) **what happens when very large or multidomain proteins are considered**, over the size range considered in its calibration, and ii) **what happens when the new ED-ENM is compared with the flexibility description obtained from submicrosecond trajectories**, where the protein is expected to display non-harmonic deformations *a priori* difficult to tackle by a pseudoharmonic- approach. We will address these questions next.

3.7.2 Influence of Protein length: Flexibility Patterns for Extremely Large Proteins

In order to answer the first question, we extended our study to several very large and multimeric proteins from MoDeL (see *Methods*), finding that ED-ENM captures well their fundamental dynamics (see *Table 3. 5 and Figure 3. 15 - Figure 3. 16*) as reported by atomistic MD simulations. This confirms that the method can be transferred to analyze large systems, difficult to tackle by MD simulations. In all cases very good results are obtained; in general, the larger the protein, the better the description of protein motions. Large proteins are more likely to contain well-defined domains and structural elements and motifs that display simpler rigid-solid motions, which are perfectly traced by our approach based on nearest-neighbours connectivity. Overall, these results confirm that the present method can be used to obtain a first hint on the dynamics of very large proteins and complexes, for which MD is prohibitive.

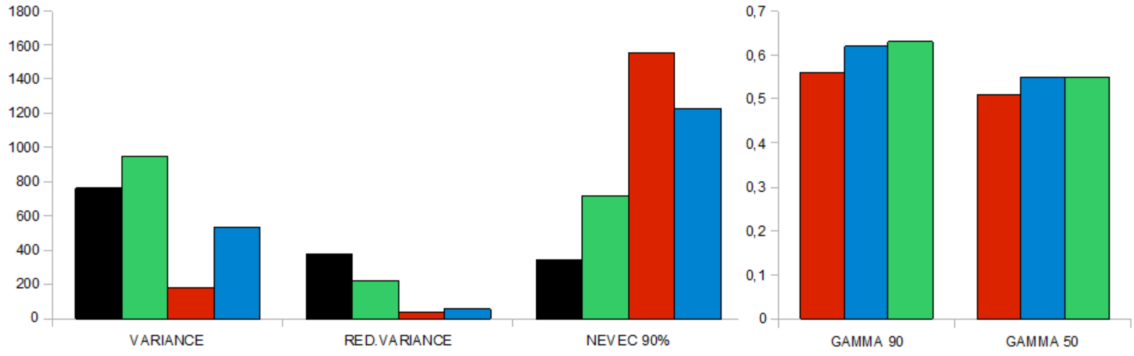


Figure 3. 15. Metrics for the comparison between MD and NMA models for extremely large proteins.

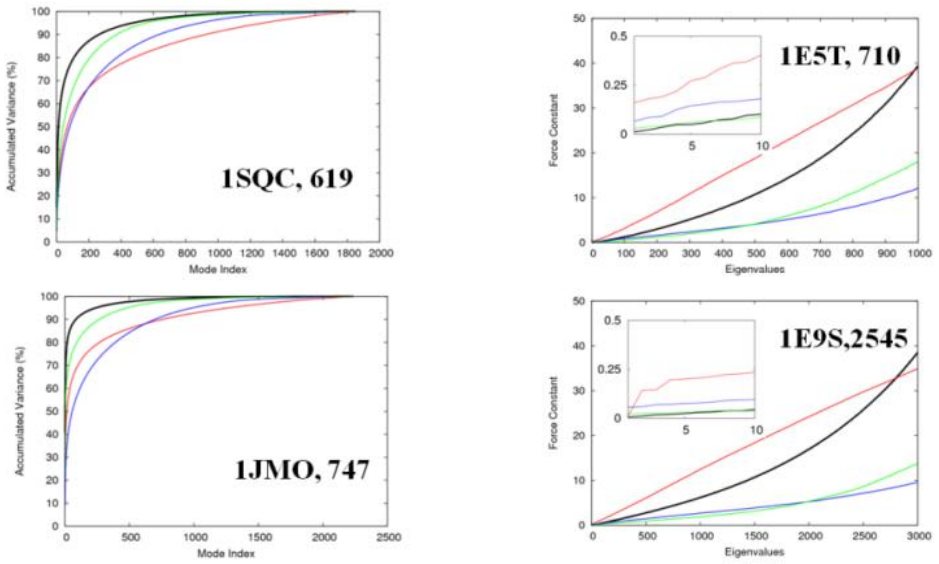


Figure 3. 16. Variance distribution and Mode Stiffness in extremely large proteins.
A) LEFT: Cumulative variance with respect to the number of eigenvectors and B) RIGHT: Strengths of the essential deformation modes (K_v, *Eq. 3.1*) for extremely large proteins (insert correspond to a zoom of first 10 eigenvalues).

Table 3. 5 Comparative Measures of Flexibility Patterns Obtained with MD and our ED-ENM NMA method for a set of very large proteins.

PDB	*Total Variance**	*Reduced Variance**	*Eigenvectors 90% Variance**	*Similarity (90% var)*[&]
1SQC (619)	306/477	109/95	264/373	0.69
1E5T (710)	485/532	226/71	217/463	0.59
1J0M (747)	1023/1131	705/621	85/267	0.62
1E9S (2545)	1237/1650	455/100	790/1753	0.62

* Values in the cells correspond to: MD (AMBER force-field)/ED-ENM method

[&] Values obtained considering standard Hess metrics (*Eq. 3.2*)

3.7.3 Influence of Simulation time length: Stability of the Modes over long trajectories

The second challenge was to compare the ED-ENM modes to those derived from very long MD trajectories (from 0.1 to 0.5-1 µs), where non-harmonic movements are likely to have more impact on the dynamics (see *Methods*). We also wanted to discard any bias introduced by the simulation length, for example, in the strong i, i+3 coupling effect observed, and to reassure that the 30 ns metatrajectories can capture the main traits of the low frequency movements. Once again, all the metrics demonstrate the robustness and generality of the ED-ENM here presented and its ability to reproduce the sub-microsecond pattern of flexibility, in particular, to correct the splitting of the soft modes observed in standard approaches (see variance distribution in *Table 3. 6* and *Figure 3.17 -Figure 3. 18*).

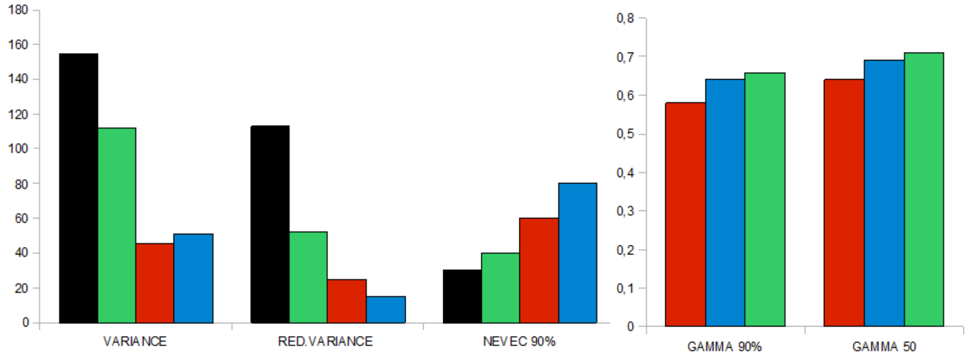


Figure 3. 17. Metrics for the comparison between MD and NMA models for long MD trajectories.
The most remarkable difference between the ED-ENM method (*green*) and the standard cutoff (*red*) and inverse (*blue*) approaches is that it concentrates most of the variance in the first dominant modes in a similar way to MD (*black*), yielding larger scale and more collective lowest-frequency motions.

Table 3. 6 Comparative Measures of Flexibility Patterns Obtained with extended MD and our ED-ENM method for four small proteins.

PDB	*Total Variance*	*Reduced Variance*	*Nevec 90%*	*Similarity (90% var)&*	*Similarity (50 eig)&*	*Lindeman's Indexes**
2GB1 (56)	43/70	30/40	23/25	0.60/0.67	0.69/0.74	0.50/0.25
1CEI (85)	300/140	230/60	17/36	0.53/0.60	0.62/0.68	0.34/0.28
1CQY (99)	43/100	21/44	59/54	0.66/0.72	0.64/0.71	0.30//0.21
1OPC (99)	230/136	170/62	21/44	0.56/0.63	0.62/0.69	0.20/0.25

* Values in the cells correspond to: MD/ED-ENM method.

& Values obtained considering standard Hess metrics; *Eq. 3.2*. Weighted similarities are typically around 0.05 to 0.1 larger.

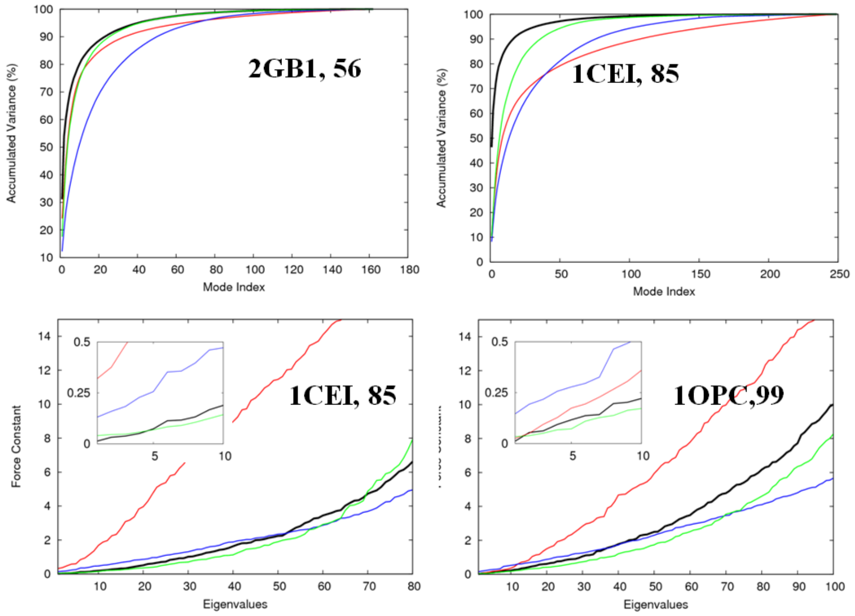


Figure 3. 18. Variance distribution in long MD trajectories.
A) TOP: Cumulative variance with respect to the number of eigenvectors in long MD trajectories. B) BOTTOM: Strengths of the essential deformation modes (K_v, *Eq. 3.1*) (insert correspond to a zoom of first 10 eigenvalues). Color code as in Fig.3.11.

However, although the variance descriptors remain in the same order of magnitude, there is a uniform tendency for all ENMs to lower similarity indexes when extending the time span of the MD (see similarity index values falling from 0.6-0.7 to 0.4-0.5 for *1cqy*, *1opc* in *Table 3. 6*). This is not surprising, since in a longer trajectory, the structures are able to explore a wider conformational subspace and thus undergo anharmonic departures from equilibrium that cannot be fully captured by any NMA-based approach as discussed above.

3.7.4 Experimental X-ray conformers and NMR ensembles

Finally, we tested the method against experimental data on flexibility from both X-ray conformers' transitions and PCA of selected NMR ensembles. First, we analyzed the ability of ED-ENM to predict functional important bound/unbound and open/close transitions between X-ray conformers. These large-scale rearrangements involve cooperative motions of domains or subunits, behaving as rigid clusters but preserving the overall fold – in this case the local cohesion prevails over inter-domain, long range interactions. Hence, a great shortcoming in continuum approaches is the over-restriction of displacements between domains, as noticed before. On the other hand, cutoff approaches display a difficult balance between violation of stereochemical constraints for lower distance thresholds, and over-restriction of motions if increased. We expected that the combination of a 6-*th* power law with a soft size-dependent cutoff, together with the strongest, inverse-square cohesion limited to neighbors, would allow more natural internal movements.

Flexibility encoded in X-ray bound-unbound pairs. First we investigated if ED-ENM can trace functional important movements upon ligand binding. Three models displaying massive transitions (rMSD between the two conformers > 5Å) were selected: i) the *1ux5→1y64* transition related to a rearrangement of helices along a flexible linker occurring upon actin binding, ii) the 1wwb→1hcf transition related to a massive tail rearrangement upon ligand binding, and iii) the *1k04→1k05* transition, in which the N-terminus and alpha-helix 1 are swapped from a highly parallel four-helical bundle following a rotation around a proline-rich hinge. Three additional models, showing local-less dramatic transitions were also studied: i) *1eia→2eia* from cubic to hexagonal tetrameric forms, a transition where the motion of the C-terminal domain (CTD) along a flexible stem allows homodimer interactions required for viral capsid assembly, ii) *1usg→1usi* transition, where a rotation around a three-stranded hinge point drives a transition from an unbound-open conformation to a bound-closed form, and iii) *1sw2→1sw5*, a conformational change related to an inter-domain rearrangement in a protein from an ABC transporter (see *Figure 3. 3* for visual description of the different transitions). Results displayed in *Table 3. 7* demonstrate that the transitions required for biological function are precisely encoded in the intrinsic deformation pattern of the ENM, as observed previously (F Tama & Sanejouand, 2001).

Table 3. 7 Overlaps between ED-ENM modes and six unbound-bound conformational transitions.
rMSD (in Å) between the bound and unbound form, overlap (in %) between deformation spaces (considering the first 5 and 10 eigenvectors) and the transition vector, and rank of maximum overlap with the bound transitions (see *Methods*).

	RMSD	Overlap (5)	Overlap (10)	Rank[&]
TRKB, 92				
1wwb↔1hcf	13.2	50	60	1(29)
FAK, 135				
1k04↔1k05	12.1	82	90	0(75)
P26, 204				
1eia↔2eia	7.7	76	80	0(71)
PROX, 265				
1sw2↔1sw5	5.0	76	80	1(51)
LBP, 319				
1usg↔1usi	7.0	91	93	0(41)
FH2, 392				
1ux5↔1y64	10	87	90	1(41)

[&] Dot product for the maximum overlap between individual eigenvectors and the transition vector in parenthesis

In all the six cases considered here the best overlap with the transition vector is found for the first 2 deformation modes, something for which random models give a probability of c.a 10^{-16}. Furthermore, extremely good overlaps are found between the transition vector and the essential deformation space of the unbound forms of the proteins: from 50 to 91% (average 77%) considering only five modes and from 60 to 93% (average 82%) if the essential space is extended to the first ten modes from the unbound form. The significance of these striking similarities becomes clear when considering that random deformations will yield to average overlaps around 0.8% (5 eigenvectors) and 1.6 % (10 eigenvectors).

Flexibility encoded in a benchmark of hinge/shear conformational transitions. To further verify our model, we extended the study to a larger benchmark including hinge, shear and complex motions from the database **MolMovDB** (Gerstein & Krebs, 1998) and compared ED-ENM results with those from standard methods. Average results for the full benchmark and detailed data for ten selected cases are displayed in *Table 3. 8:* there are four structures undergoing large transitions (RMSd > 7 Å) and six more with local, less dramatic changes (RMSd 2-6 Å). However, the ED-ENM provides in most cases the best agreement between the transition vector and the harmonic deformation space. In the open forms, considering only the five modes, the overlaps range from 0.60 to 0.95 (average 0.7), and from 0.70 to 0.97 (average 0.76) considering ten modes (see *Table 2*); note that random deformations would yield overlaps around 0.08 (5 eigenvectors) and 0.16 (10 eigenvectors).

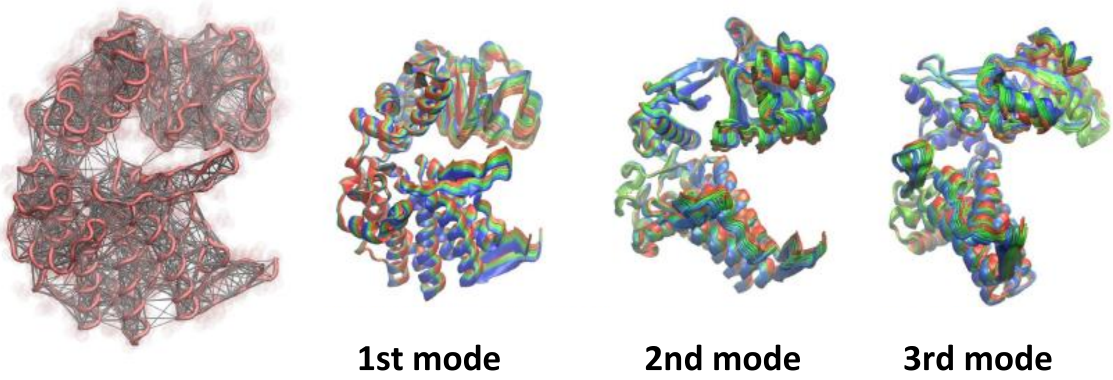


Figure 3. 19. Lowest frequency modes of the open conformer of the Oligopeptide binding protein, OppA (*1rkm*).

The transition to the close bound state is fully mapped in a 90% by the first ED-ENM mode of the open state (*right*).

There is a systematic trend to better performance of the ED-ENM (2-5%) regardless of the extent of the transition, especially considering only the first five modes. The greatest improvement is achieved for the closed forms, more difficult to treat since the can be easily over-constrained by long-range springs: in this case the $\gamma_{(5)}$ increases by nearly 10% (from 0.50 in standard ENM to 0.60). The agreement is remarkable in the most challenging cases, where other ENMs fail dramatically (see for example, the *closed → open* transition for *1ckmB*, *1ama*, *1dap*). These notable differences are related to the concentration of the conformational change in the first dominant eigenvectors. Accordingly, the rank differences are also smaller and the best overlapped eigenvectors closest to the transition direction. In summary, the ED-ENM displays a higher cooperativity and less dispersion of the motions - as in the above comparison with ED - and thus traces the functional changes with fewer modes.

Table 3. 8 Overlaps for the NMAFIT 54 conformational transitions benchmark.
RMSd (in Å) between X-ray conformations, overlaps (in %) between essential deformation spaces (considering the first 5 and 10 eigenvectors) and the transition vector, and Rank of maximum overlap (see *Methods* for description of the different metrics) for the cutoff, the inverse exponential model and the ED-ENM.

Length (CATH)	PDB	RMSD	*Overlap(5)**	*Overlap(10)**	Rank & *Overlap$_{max}$* *&
101	**1l5e (open)**	8.8	0.76 / 0.43 / 0.81	0.81 / 0.76 / 0.85	0 (**0.70**) / 0 (0.66) / 0 (0.65)
	1l5b (closed)		0.83 / 0.80 / 0.81	0.86 / 0.85 / 0.87	1 (0.27) / 1 (0.28) / 2 (**0.55**)
148	**1cfd (open)**	10.2	0.88 / 0.93 / 0.94	0.93 / 0.94 / 0.95	1 (0.38) / 0 (**0.62**) / 1 (0.55)
	1cfc (closed)		0.83 / 0.89 / 0.89	0.93 / 0.92 / 0.94	1 (**0.55**) / 0 (0.45) / 1 (**0.55**)
214	**4ake (open)**	8.3	0.90 / 0.90 / 0.92	0.93 / 0.92 / 0.93	0 (**0.67**) / 0 (0.38) / 0 (**0.67**)
	1ake (closed)		0.55 / 0.57 / 0.64	0.61 / 0.68 / 0.71	0 (0.32) / 0 (0.36) / 0 (**0.40**)
219	**1nbv (H) (open)**	2.2	0.69 / 0.69 / 0.70	0.73 / 0.72 / 0.73	2 (**0.68**) / 0 (0.29) / 0 (0.32)
	1cbv (H) (closed)		0.68 / 0.69 / 0.71	0.72 / 0.71 / 0.72	2 (0.38) / 2 (**0.40**) / 0 (0.37)
271	**1urp (open)**	7.7	0.96 / 0.93 / 0.95	0.96 / 0.95 / 0.97	1 (**0.94**) / 1 (0.72) / 1 (0.80)
	2dri (closed)		0.83 / 0.82 / 0.88	0.86 / 0.88 / 0.92	0 (0.62) / 0 (0.56) / 1 (**0.71**)
317	**1ckm (A) (open)**	4.3	0.93 / 0.91 / 0.93	0.94 / 0.93 / 0.95	0 (0.86) / 0 (0.44) / 0 (**0.88**)
	1ckm (B) (closed)		0.21 / 0.49 / 0.57	0.65 / 0.73 / 0.78	6 (**0.29**) / 2 (0.14) / 4 (0.23)
320	**3dap (open)**	5.8	0.89 / 0.90 / 0.93	0.94 / 0.92 / 0.95	0 (**0.75**) / 1 (0.58) / 0 (0.68)
	1dap (closed)		0.20 / 0.18 / 0.27	0.44 / 0.62 / 0.78	9 (0.19) / 7 (0.33) / 4 (0.22)
401	**9aat (open)**	2.2	0.15 / 0.07 / 0.55	0.68 / 0.64 / 0.71	5 (0.26) / 5 (**0.45**) / 4 (0.44)
	1ama (closed)		0.07 / 0.08 / 0.60	0.68 / 0.67 / 0.76	6 (0.30) / 5 (**0.39**) / 6 (0.30)
452	**1bnc (open)**	5.4	0.84 / 0.85 / 0.87	0.87 / 0.90 / 0.90	0 (**0.83**) / 0 (0.71) / 0 (0.81)
	1dv2 (closed)		0.70 / 0.69 / 0.76	0.77 / 0.80 / 0.85	4 (0.22) / 0 (0.40) / 0 (**0.48**)
517	**1rkm (open)**	5.8	0.93 / 0.92 / 0.93	0.94 / 0.94 / 0.95	0 (0.91) / 0 (0.84) / 0 (**0.92**)
	2rkm (closed)		0.62 / 0.64 / 0.67	0.68 / 0.75 / 0.73	1 (0.32) / 0 (**0.52**) / 0 (0.42)
NMAFIT Average		**open**	**0.66 / 0.67 / 0.70**	**0.75 / 0.75 / 0.76**	0.9(0.69)/ 0.6(0.57)/0.6(0.67)
		closed	**0.51 / 0.53 / 0.58**	**0.65 / 0.67 / 0.69**	3.0(0.35)/1.7(0.38)/ 1.7(0.42)

[*] Values in these columns as cutoff/inverse/ED-ENM [&] In parenthesis, dot product of the best overlapped vector

Flexibility encoded in NMR ensembles. Finally, we analyzed the ability of ED-ENM to approach the structural diversity of NMR ensembles (*Figure 3. 20*). The PCA of high-quality NMR ensembles has been shown to correlate well with ED and ENMs normal modes (Abseher, Horstink, Hilbers, & Nilges, 1998; L.-W. Yang, Eyal, Bahar, & Kitao, 2009b), and thus it is expected to provide qualitative information on the more flexible regions of a protein.

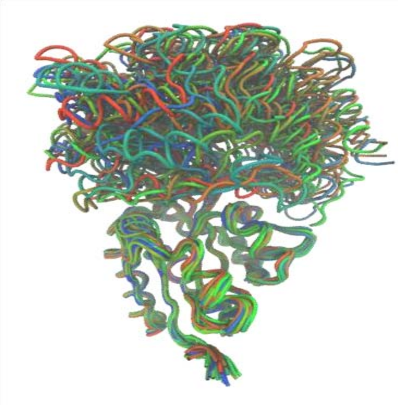


Figure 3. 20. NMR ensemble (*2d21*).

Table 3. 9 Comparative Measurements of Flexibility Patterns obtained from NMR and Normal Modes.

PDB	N	M	*Overlap(5)**			*Overlap(10)**			*Overlap$_{max}$*		
1ro4	**58**	**35**	0.56	0.52	0.58	0.57	0.53	0.62	0.59	0.33	**0.72**
1e9t	**59**	**59**	0.51	0.48	0.63	0.53	0.54	0.58	0.48	0.57	**0.64**
1bw5	**66**	**50**	0.69	0.59	0.74	0.56	0.56	0.62	0.53	0.63	**0.78**
2eot	**74**	**32**	0.52	0.42	0.61	0.56	0.41	0.55	0.57	**0.76**	0.54
1a6x	**87**	**49**	0.55	0.44	0.56	0.41	0.37	0.48	0.54	0.37	**0.95**
1bve	**99**	**28**	0.47	0.52	0.49	0.33	0.36	0.37	0.81	0.7	**0.88**
1q06	**101**	**55**	0.57	0.58	0.57	0.51	0.60	0.57	0.51	0.58	**0.77**
2czn	**103**	**38**	0.62	0.55	0.66	0.50	0.53	0.51	**0.87**	0.65	0.70
1a90	**108**	**31**	0.36	0.44	0.49	0.38	0.40	0.42	0.65	0.49	**0.83**
2bo5	**120**	**44**	0.54	0.37	0.58	0.55	0.45	0.56	**0.73**	0.57	0.68
1e5g	**120**	**50**	0.71	0.70	0.69	0.60	0.64	0.63	0.93	0.90	**0.96**
1cmo	**127**	**43**	0.70	0.61	0.70	0.56	0.52	0.59	0.56	**0.60**	0.52
1iti	**133**	**31**	0.53	0.65	0.59	0.46	0.51	0.44	0.78	0.78	**0.89**
1c89	**134**	**40**	0.70	0.76	0.80	0.55	0.60	0.63	0.53	**0.69**	0.63
1xsb	**153**	**39**	0.46	0.43	0.49	0.44	0.38	0.43	0.89	0.41	**0.92**
1bf8	**205**	**20**	0.47	0.55	0.54	0.43	0.47	0.48	0.86	0.78	**0.88**
1by1	**209**	**20**	0.55	0.55	0.56	0.42	0.46	0.47	0.50	**0.65**	0.53
1n6u	**212**	**22**	0.63	0.61	0.59	0.58	0.58	0.58	**0.64**	0.37	0.60
2jz4	**299**	**20**	0.53	0.51	0.61	0.41	0.40	0.45	0.49	**0.80**	0.74
2d21	**370**	**20**	0.60	0.56	0.65	0.50	0.47	0.50	**0.67**	0.62	0.62
Average			0.56	0.54	0.61	0.49	0.49	0.53	**0.53**	0.61	0.74

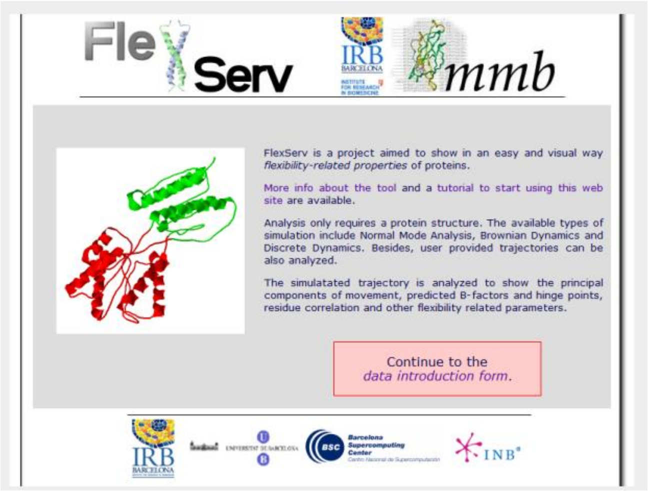
[*] Values in these columns as cutoff/inverse/ED-ENM (see *Methods* for metrics definitions)

The analysis of 26 selected NMR multiple structures shows striking correlations with the first ENMs modes (see *Table 3. 9*), in agreement with the above mentioned observations, and supports the validity of ENMs to sample the near-equilibrium conformational space in solution. In this case, it is also clear that the ED-ENM method outperforms the cutoff method and the Kovacs approach, especially when considering only the first 5 eigenvectors whose overlap raises from 0.56/0.54 to 0.61 (an average increase of around 5%) and the best overlapped pair, which increases from a 0.65/0.61 average to 0.74 (more than a 10% increase), reaching values near 0.90 (see *1bve*, *1iti*, *1bf8*) or even above (*1a6x*, *1e5g* and *1xsb*). In more than half of the proteins (11 cases), the best overlapped vector is found in the ED-ENM method, followed by the inverse (5 cases) and cutoff (4 cases) approaches, following the trend observed in the rest of tests against experimental flexibility sources.

3.7 The FlexServ interface

The great computational efficiency of the ED-ENM method allows its implementation in webservers. We have developed an integrated platform to launch coarse-grained simulations, **FlexServ:** http://mmb.pcb.ub.es/FlexServ, which provides a friendly access for non-expert users in an intuitive graphical interface. FlexServ (Camps et al., 2009) is a web-based tool for the analysis of protein flexibility. The server incorporates powerful protocols for the three standard coarse-grained simulation methods presented in Chapter 2: ***Normal Mode Analysis*** **(NMA),** ***Brownian dynamics*** **(BD)** and ***Discrete Molecular Dynamics*** **(DMD)**, as well as for PCA analysis of atomistic trajectories (ED). Besides allowing to launch coarse-grained simulations, and to analyze the results, it also incorporates direct links with the main structural databases.

The NMA interface allows the calculation of the normal modes in the cutoff-like and Kovacs schemes as well as using the ED-ENM method. The server allows a complete analysis of flexibility using a large variety of metrics, including most of the measures introduced in this Chapter: from basic geometrical analysis, B-factors, essential dynamics, stiffness analysis, to collectivity measures, Lindemann's indexes, residue correlation, chain-correlations, dynamic domain determination, hinge point detections, etc. The resulting data is presented through a web interface as plain text, and intuitive 2D and 3D graphics.



FlexServ is a project aimed to show in an easy and visual way *flexibility-related properties* of proteins.

More info about the tool and a tutorial to start using this web site are available.

Analysis only requires a protein structure. The available types of simulation include Normal Mode Analysis, Brownian Dynamics and Discrete Dynamics. Besides, user provided trajectories can be also analyzed.

The simulatated trajectory is analyzed to show the principal components of movement, predicted B-factors and hinge points, residue correlation and other flexibility related parameters.

Figure 3. 21. The FlexServ website
Screen capture of the FlexServ webpage.

3.8 Summary

The ability of the elastic network NMA models to predict the intrinsic motions of proteins has been widely demonstrated in the last years. However, predictions are eminently qualitative and often need *ad-hoc* re-scaling of the mode amplitudes. In comparisons with MD, ENMs yield a sparser pattern of flexibility, related to their harmonic character, and the information required for a realistic description of a functional motion is dispersed into multiple modes. Another problem has been the lack of consensus in the ENMs refinement, mainly due to the scarcity of measurements of protein flexibility. In this work we have used atomistic simulations as reference to infer more realistic connectivity rules and obtain a scaling of the force constants.

Our findings point out that, whereas the directions of the low frequency motions depend entirely on the network topology, the strength of the constants determines the amplitude of the modes and the variance profiles. We found that low frequency motions are mainly encoded on the strong, covalent interactions linking sequentially close residues, with the long range noncovalent interactions playing a minor tuning role. These constraints led to the formulation of an **ED-refined ENM (ED-ENM)**, based on a simple hybrid potential based on chain topology. The method proposes a simple and robust scaling of the local backbone and long-range contacts. The new method has been validated with the largest available collection of consensus force-field meta-trajectories and has shown ability to improve the representation of local and global flexibility from small to large proteins.

The goal was not to reproduce any particular flexibility measurement (such as B-Factor profiles), but rather to develop a general method able to trace motions better than or at least as well as the best-performing standard approach, for the widest range of descriptors and protein sizes and folds. Clearly, when considering all the flexibility measurements here presented, the ED-ENM outperforms standard approaches without any *ad-hoc* adjustments.

Finally, we have demonstrated that ED-ENM captures the experimental flexibility, such as large functional X-ray transitions, or NMR structural ensembles, with remarkable precision. Therefore, the results demonstrate that ED- ENM can be a useful alternative to well established coarse-grained NMA methods. In a wider context, the ability of a simple nearest-neighbors model to match MD and to trace functional changes suggests that local covalent topology greatly influences the intrinsic motions of proteins, and that this shape-encoded deformability guides biological conformational changes.

3.10 Publications from this chapter

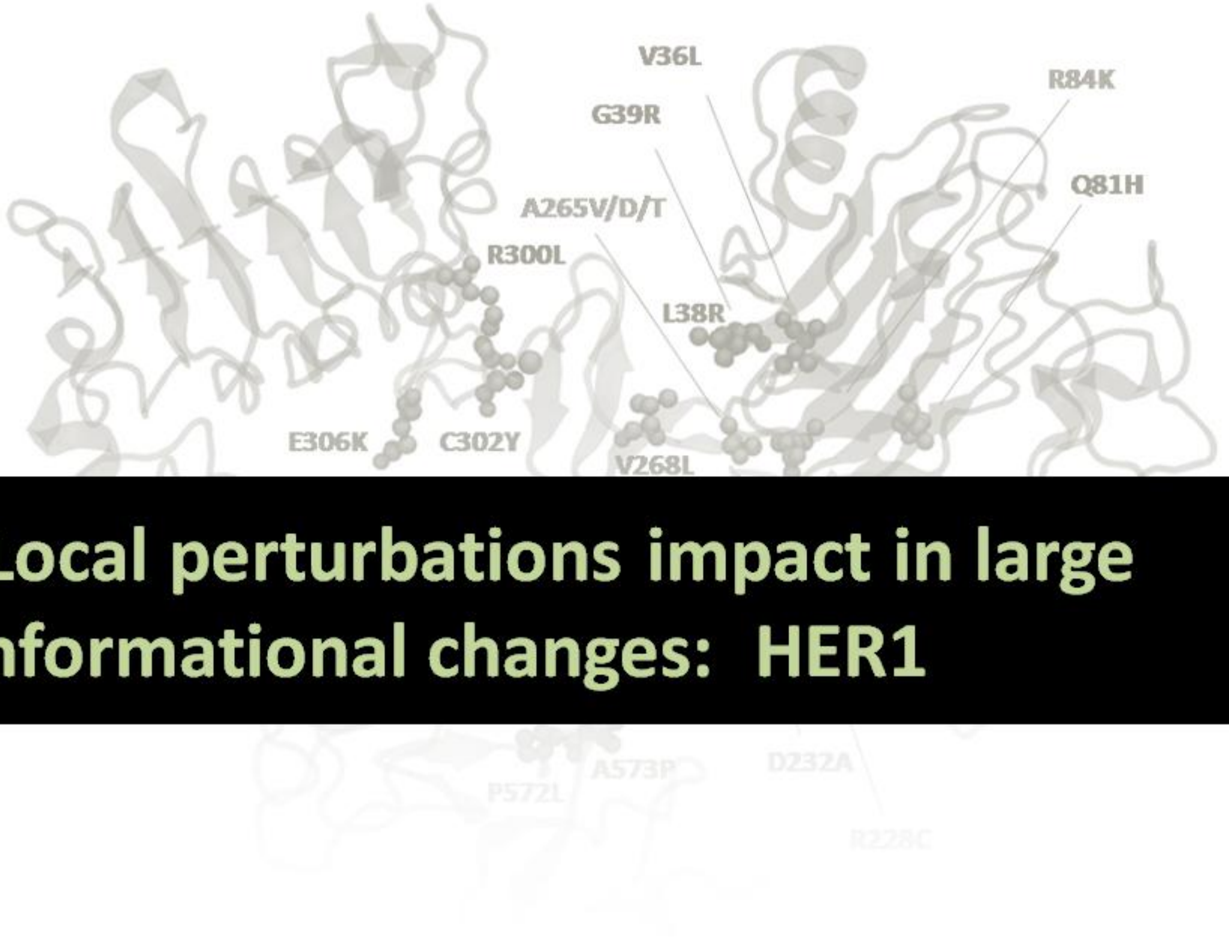
- Camps J., Carrillo O., Emperador A., Orellana L., Hospital A., Rueda M., Gelpi JL. & Orozco M. ***"FlexServ: An integrative tool for the Analysis of protein Flexibility"*** *Bioinformatics* (2009) 25 (1): p1709–1710.

- Orellana L., Rueda M., Ferrer-Costa C., López-Blanco J.R., Chacón P. and Orozco M. "***Approaching Elastic Network Models to Molecular Dynamics Flexibility"*** *J. Chem. Theory Comput.* (2010) 6: p2910–2923.

  Featured Cover Paper: JCTC September 2010 Cover

“Owing to this struggle for life, any variation, however slight and from whatever cause proceeding, if it be in any degree profitable to an individual of any species… will tend to the preservation of that individual, and will be inherited by its offspring.”

Charles Darwin



4- Local perturbations impact in large conformational changes: HER1

This page intentionally left blank

Chapter 4 Local Perturbations impact in large-scale conformational changes: the HER1 case

In this central part of the present thesis, we take one step further and want to explore which are the mechanisms behind large-scale protein flexibility responsible for function. *How a protein shape and its movement are related, and how they change and evolve through mutations? How can a single residue change cause protein malfunction and disease? Which are the key regions orchestrating protein dynamics?* To answer these questions, we focus on a protein undergoing a large-scale conformational change clearly linked to function, the **Epidermal Growth Factor Receptor (EGFR)**. The EGFR family is the prototype of protein tyrosine kinase receptors, which are key regulators of cellular growth. EGFR signaling is thought to be initiated by a dramatic ligand-favored conformational change of the extracellular domain (*ectodomain* or sEGFR) from a closed, self-inhibited tethered monomer, to an open untethered state, which exposes a loop required for dimerization and activation. Many evidences suggest that the ectodomain also untethers spontaneously, and it is widely assumed that unbound untethered states are open as in crystal structures. In the brain tumor *glioblastoma multiforme*, multiple clusters of point missense mutations appear located in the extracellular domain, which cause ligand-independent activation, and presumably, an active, open conformation. Here we explore impact of cancer mutation clusters on the ectodomain dynamics using a perturbative screening based on the *ED-ENM* algorithm (**Chapter 3**) to locate dynamical hot spots for the conformational changes, determine the impact of mutations and guide Molecular Dynamics simulations comprising several microseconds. Our results show that key interdomain contacts crucial for conformational changes are precisely mutated in cancer, suggesting a novel oncogenic mechanism based on the perturbation of the native motions. Simulation of a mutation targeting a dynamical hot spot reveals the spontaneous formation of an untethered intermediate, which is unexpectedly closed but capable to dimerize and which exposes the mAb806 conformational epitope characteristic of tumors. Our findings point out to different mechanisms driving ligand-dependent and independent untethering, and highlight the connection between oncogenic mutations and the ectodomain conformational changes.

4.1 Mutation, protein evolution and disease

How is it possible that a change in a single residue perturbs the function of a protein? This question is of great relevance, since it underlies the basis of many diseases, but is also the driving force for protein evolution. Although function and dynamics are robustly encoded in the structure, perturbations in sensitive regions can have a dramatic impact. Near-equilibrium equilibrium fluctuations allow proteins to explore the conformational space encoded in their structure. Back in the 60s, the **Monod-Wyman-Changeux (MWC)** (Monod et al., 1965) model of allostery proposed conformational selection amongst accessible conformers as the basis for molecular recognition: according to this model, the presence of a ligand just shifts the population in favor of the conformer best suited to bind, in contrast to the induced-fit model where the ligand 'induces' the change in conformation (Koshland, 1958). However, the concept of allostery has evolved to a more complex vision where multiple dynamic ensembles coexist and change in response to perturbations (Acuner Ozbabacan, Gursoy, Keskin, & Nussinov, 2010; Csermely, Palotai, & Nussinov, 2010; del Sol, Tsai, Ma, & Nussinov, 2009). Understanding which regions control the intrinsic dynamics of a protein and how different states are sampled can help to unravel the mechanisms by which mutations change biological function and drive evolution.

4.2 The HER case: a large-scale mutation-sensitive transition

As presented in the introduction, a dramatic example of point mutations related to changes in activity is the *Epidermal Growth Factor Receptor* main isoform, EGFR or HER1 (Human EGFR 1). HER1 is the founding member of the large family of transmembrane tyrosine kinase receptors (TKRs), which are fundamental regulators of cell growth, survival, differentiation and migration in mammalian cells (Schneider & Wolf, 2009; Wieduwilt & Moasser, 2008). All proteins of the TKRs family act as sophisticated molecular machines, transducing information in a unidirectional fashion from the environment across the cell membrane. These glycoproteins consist of a large extracellular region followed by a single transmembrane-spanning region and a cytoplasmic kinase domain, acting in a highly concerted fashion. In humans the EGFR family includes four receptors: EGFR itself (HER1, ErbB1), ErbB2 (HER2/Neu), ErbB3 (HER3) and ErbB4 (HER4), which play a crucial function in both ectodermal and mesodermal cell lineages, being thus very potent oncogenes: disregulation of their signaling pathways by overexpression or hyperactivation has been related to many carcinomas such as erythroid leukemia, fibrosarcoma, angiosarcoma, glioblastoma, and melanoma (Baselga, 2002; Zandi, Larsen, Andersen, Stockhausen, & Poulsen, 2007). The ligand-*sensing* piece of these receptors is the 600 residue long extracellular domain (sEGFR), composed of a tandem repeat of two homologous ligand binding *large* subdomains (I and III) of around 200 residues length, plus two *small* or cysteine-rich subdomains (II and IV) of around 130 residues length, containing serial Cys-linked modules. In the inactive states crystallized, a self-inhibitory *tether* in domain IV (T561-K585) holds hidden the *dimerization arm* in domain II (T246-M253).

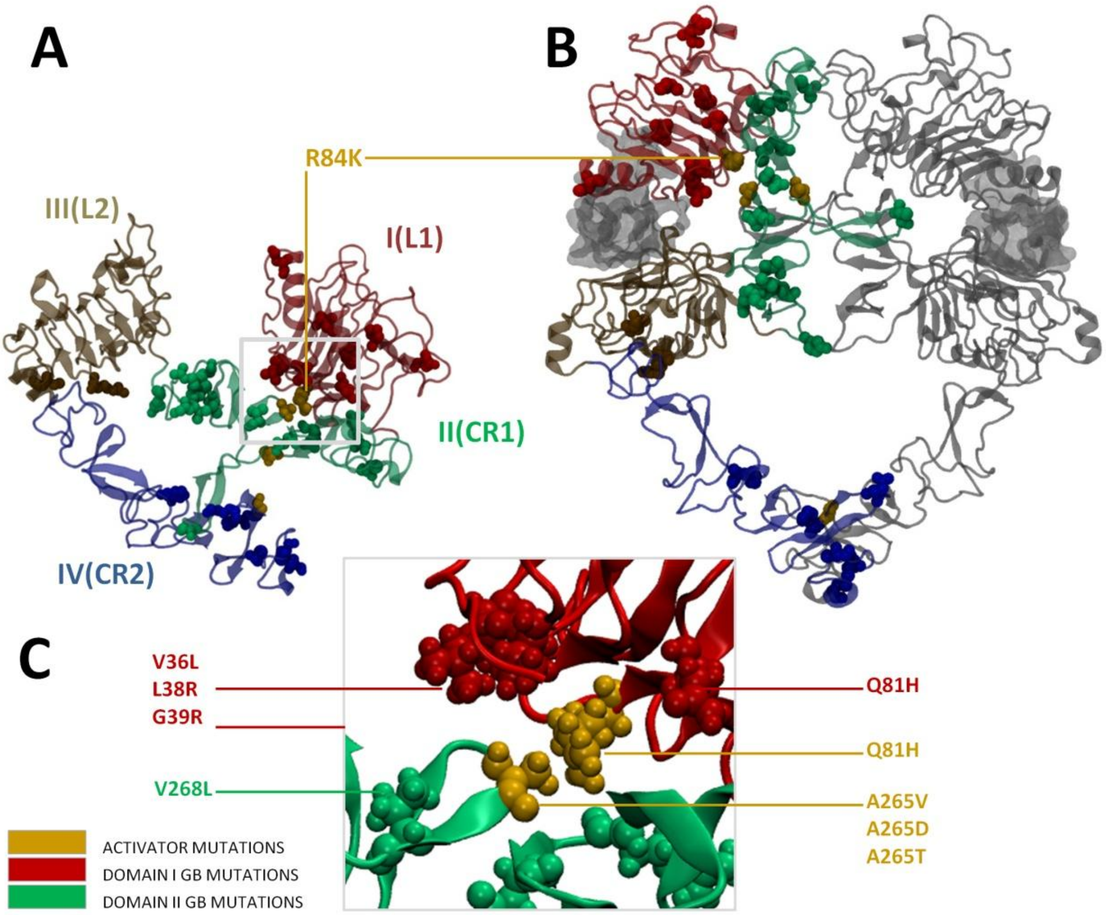


Figure 4. 1 **sEGFR inactive closed monomer (A) and active open dimer (B).**

Known mutations in *glioblastoma* are shown*,* highlighting in yellow those reported to increase activity. Note the evident clustering of mutations at the interface between domain I (red) and II (green), and domains II and IV (blue), as well as the position of R84K mutation away from domain I-III binding cleft as well as domain II dimerization interface. Mutations in domain III (ochre) - IV interface are also described in other tumors. (C) A close-up to domain I-II interface where most *glioblastoma* mutations appear clustered.

Despite the effort focused in its characterization, the molecular mechanism for activation at the extracellular level is still not fully understood, and the question whether the ligand triggers a tethered to untethered transition or whether there is a preexisting conformational equilibrium remains unanswered. Crystal structures suggest that ligand binding favors a large conformational change, from the tethered closed inactive monomer (Ferguson et al., 2003) to an untethered, open dimer (Garrett et al., 2008; Ogiso et al., 2002) thought to represent the active state (*Figure 4. 1, A* and *B*, respectively). In this process, the subdomains I and III reorient to bind simultaneously EGF, whereas the *tether* releases the *dimerization arm* to form key receptor-receptor contacts in the active dimer. Recent long simulations of HER4 (Du et al., 2012) have confirmed this classical induced-fit model where ligand binding drives the transition from the closed to an extended, open-like conformation with a free dimerization arm. There are however several observations that do not fit in the ligand-driven dimerization scheme. For example, although the crystallographic dimer is doubly-ligated, experimental evidences point out that unbound and singly-ligated (Liu et al., 2012) dimeric species also exist suggesting that untethering also occurs in the absence of ligand. Therefore, current thinking rather assumes a conformational selection mechanism, where the ectodomain samples untethered conformations, which are stabilized by ligand binding. It is widely assumed that these unbound untethered monomers would be open as in the crystal dimer. Although there is data supporting a preexisting tethered-untethered equilibrium (Kozer, Rothacker, Burgess, Nice, & Clayton, 2011), a spontaneous transition of unbound sEGFR towards an extended form is still to be detected experimentally (Dawson, Bu, & Lemmon, 2007) or in microsecond-long MD (Arkhipov et al., 2013). Intriguingly, one of the features of the *tethered →untethered* conformational change seems to be the transient exposure of a Cys-bonded module (C287-C302) in domain II (Johns et al., 2004). This region acts as “cryptic” epitope triggering immune response by antibodies mAb806 and mAb175 against overexpressed and N-terminally truncated EGFR (EGFRvIII), but appears hidden in all known ectodomain structures (Garrett et al., 2009).

In comparison with the thorough studies on kinase-targeting mutations (Shan et al., 2012; Shan, Arkhipov, Kim, Pan, & Shaw, 2013; Sutto & Gervasio, 2013), missense mutations mapping on the EGFR extracellular domain (found in brain gliomas) have received little attention to date. These *gain-of-function* mutations (see *Figure 4. 1, C* and *Table 4. 1*) cause ligand-independent activation of EGFR (Huang, Xu, & White, 2009; Lee et al., 2006) by an unknown mechanism: although some target the self-inhibitory tether or the ligand-binding sites, most of them appear far away, clustered at interdomain conserved surfaces. In order to figure out *if* and *how* these mutations alter the ectodomain conformational change, we performed a comprehensive study of the dynamics of Wild Type (WT) and mutated forms of sEGFR. By combining ENMs (Atilgan et al., 2001; Tirion, 1996) with near 5 microseconds of atomistic MD (Karplus & McCammon, 2002), we have characterized the major conformational states of HER1, outlining a complete picture of spontaneous and ligand-induced untethering. Simulations of ectodomain activator mutations have revealed a spontaneous untethering process, surprisingly, not to an extended but to a close untethered state which may correspond with the immunogenic mAb806- intermediate in cancer. Our results point to an unexpected mechanism for ligand-independent activation, and provide a simple molecular explanation for oncogenic mutations mapping on sEGFR.

Table 4. 1 Summary of isolated and clustered missense mutation in the EGFR ectodomain targeting interdomain regions, as registered in the COSMIC database.

Mutated residues are considered as neighbors up to three residues apart in the sequence. Residues related to increased activity underlined in column 1.

Mutation	Interdomain region	Sequence Cluster	Tissue & Frequency	Conservation (ConSurf)	Kinase activity
V36R	I-II	3X-cluster	Glioma, 1	**8**	
L38R/P			Glioma, 2 & Lung, 3 /Colon,1	**8**	
G39R/W			Glioma, 1 & Colon, 2/Colon,1	**9**	
Q81H		8X-cluster	Glioma, 2	**7**	
R84K/G			Glioma, 9 & Skin, 1 / Colon, 1	**9**	Ligand-independent kinase, Tumorigenic (Lee *et al.*)
C195G		19X-cluster	Glioma, 1	8	
R198C			Glioma, 2 & Colon carcinoma, 1	6	
R228C		23X-cluster	Glioma, 1	6	
D232A			Glioma, 1	4	
T239P			Glioma, 4	3	Ligand-independent kinase, Tumorigenic (Lee *et al.*)
T249N			CNS Glioma, 1	4	
A265V/D/T		26X-cluster	Glioma, 20/3/3	5	Ligand-independent kinase, Tumorigenic (Lee *et al.*)
V268L			Glioma, 1	**7**	
G298C	II-III	30X-cluster	Thyroid carcinoma, 1	4	
V299I			Lymphoid neoplasm, 20/3/3	2	
R300L			Glioma, 2	7	
C302Y			Glioma, 1	9	
E306K*			Glioma, 1	1	
R427F	III-IV		Lung, 1		
I451V			Upper dig. tract, 1		
P572L	II-IV	57X-cluster	Glioma,2/ Colon, 1	8	Ligand-independent kinase, Tumorigenic (Lee *et al.*)
A573P/T			Glioma, 1/ Lung, 1	1	
G574V			Glioma, 16	6	
E578Q			Glioma, 1	3	

4.3 Methods

Structural data and naming scheme for extracellular domains of HER. We follow the naming scheme for sEGFR, distinguishing four CATH (Orengo et al., 1997) subdomains: large domain I (residues 1-190), cysteine-rich domain II (191-309), large domain III (310-481) and cysteine-rich domain IV (482-614). Our reference are the coordinates of the 612-long fragment Glu3-Thr614 from closed (PDB: *1nql*) and open dimeric (PDB: *3njp, chain A*) structures.

***Elastic Network-Normal Mode Analysis*.** We use the ED-ENM algorithm (Orellana et al., 2010) based on the *Anisotropic Network Model* (ANM) (Atilgan et al., 2001; Tirion, 1996) (see details in **Chapter 3**).

Perturbation screening of interdomain contacts. Structural perturbation methods for ENMs are used to identify the residues coupled with important motions and whose mutations may impact the protein's function (Zheng Brooks, 2005; Zheng et al., 2007, 2005; Matsumoto el al., 2008). Usually, an energetic perturbation dE_i is introduced at residue *i* as a change in the force constants of those springs connecting residue i to its neighbors (residue j) within a cutoff, and the response of the rest or a subset of residues *j*, measured in terms of the changes in the mean-square fluctuations. The residues *i* causing the largest change in the fluctuations of residues *j* are considered dynamically important residues. Due to the robustness of ENMs and the highly nonlinear effect of the force constants, it is difficult to model realistically the effects of perturbative changes in the volume or charge of a residue.

Here we design a novel technique, not residue-centered but region-centered, to trace those interactions whose perturbation is expected to have the largest impact in the essential motions. The idea is very simple: to mimic the effects of mutations in the elastic network, we assume that **a pathogenic change will disrupt locally *all the pattern of neighbor residue interactions***, impacting not one single spring, but a bunch of them close in the space. Thus, we screen pair interactions by perturbing one by one each spring plus its local neighbors. For each perturbed state we compute the essential movements using ENM, and the effect is measured by the similarity between the modes and the transition vector, which recapitulates the known movements. Looking at the perturbations impacting more dramatically the alignment between the modes and the conformational change, we detect dynamically hot regions. For computational efficiency, we focus our screening in inter-domain contacts (pairs of residues which belong to different CATH domains and are located at less than 11Å in the close (*1nql*) state). This yields a list of 439 contacts, which are used as seeds to detect small regions (2, 3 and 4 Å cutoffs radii), which define neighboring interactions within a threshold (the distance between two springs is defined by the mean point between their centers). Then, for each primary spring *ij*, we have from 1-5 (2Å threshold) to 10-30 (4Å threshold) non-sequential neighbor springs. Altogether, each primary interaction and all its neighboring contacts are perturbed by either decreasing (leading the spring to zero) or increasing (to 10 kcal/mol.Å^2, which is above the network background) their strength. Therefore, we scan each primary interaction between a pair of residues (*i, j*) performing *S*=6 different perturbation experiments (the result of increasing and decreasing the strength of the local contacts for 3 different neighbors definitions). Note that the total number of springs is around 33000 in the ED-ENM, so our perturbation affects to less than 0.1% of the system. Once a locally perturbed elastic network is defined, we compute the modes. For each perturbation *i, j* and each mode k we compute then the resulting overlap with the vector mapping the close -> open transition, α_{ij}^{S} (see below, *Eq. 4.4)*; this overlap computed over all the interactions defines a ***"baseline"*** (an average overlap, ***Avg*** $\boldsymbol{(\alpha_k^S)}$) that represents the dynamical effect of a random perturbation (*i.e.* the background).

Sensitive interactions would largely depart from the baseline (***"signal peaks"***), and thus could be evaluated by the number of times above the average standard deviation in the overlap, ***nSD*** $\boldsymbol{(\alpha_k^S)}$***:***

$$nSD\ (\alpha_{ij}^{S}) = \frac{\left|\alpha_{ij}^{S} - Avg(\alpha_{k}^{S})\right|}{SD(\alpha_{k}^{S})}$$

(Equation 4. 1 *Overlap Average SD*)

Then, for each essential mode k and each perturbation *ij*, we add the nSD accumulated across the *S*=6 perturbation experiments:

$$\zeta_{ij}^{k} = \sum_{1}^{S} nSD\ (\alpha_{ij}^{S})$$

(Equation 4. 2 *Accumulated Average SD*)

Then we normalize this sum, ζ_k, for all the perturbations (in a scale from 0 to 100), to obtain a final score for the mode *k,* ζ_k*'*. The process is repeated for all the first *m*=5 important modes and finally, the normalized results for each mode are added to derive a simple perturbation **dynamical impact score** for the *m*-essential motion space, ζ_m.:

$$\zeta_{ij}^{m} = \sum^{m} \zeta_{ij}^{k}\,'$$

(Equation 4. 3 *Dynamical Impact Score*)

The maximal score here is 400, since modes 1st -2nd are considered together due to their close frequencies. A high score reflects a SD for an interaction *systematically* over the baseline across different perturbative experiments. Here most interactions have scores below 100-150; thus scores over 150-200 for a set of neighbor interactions identify a region where a mutation impacts the essential deformation modes.

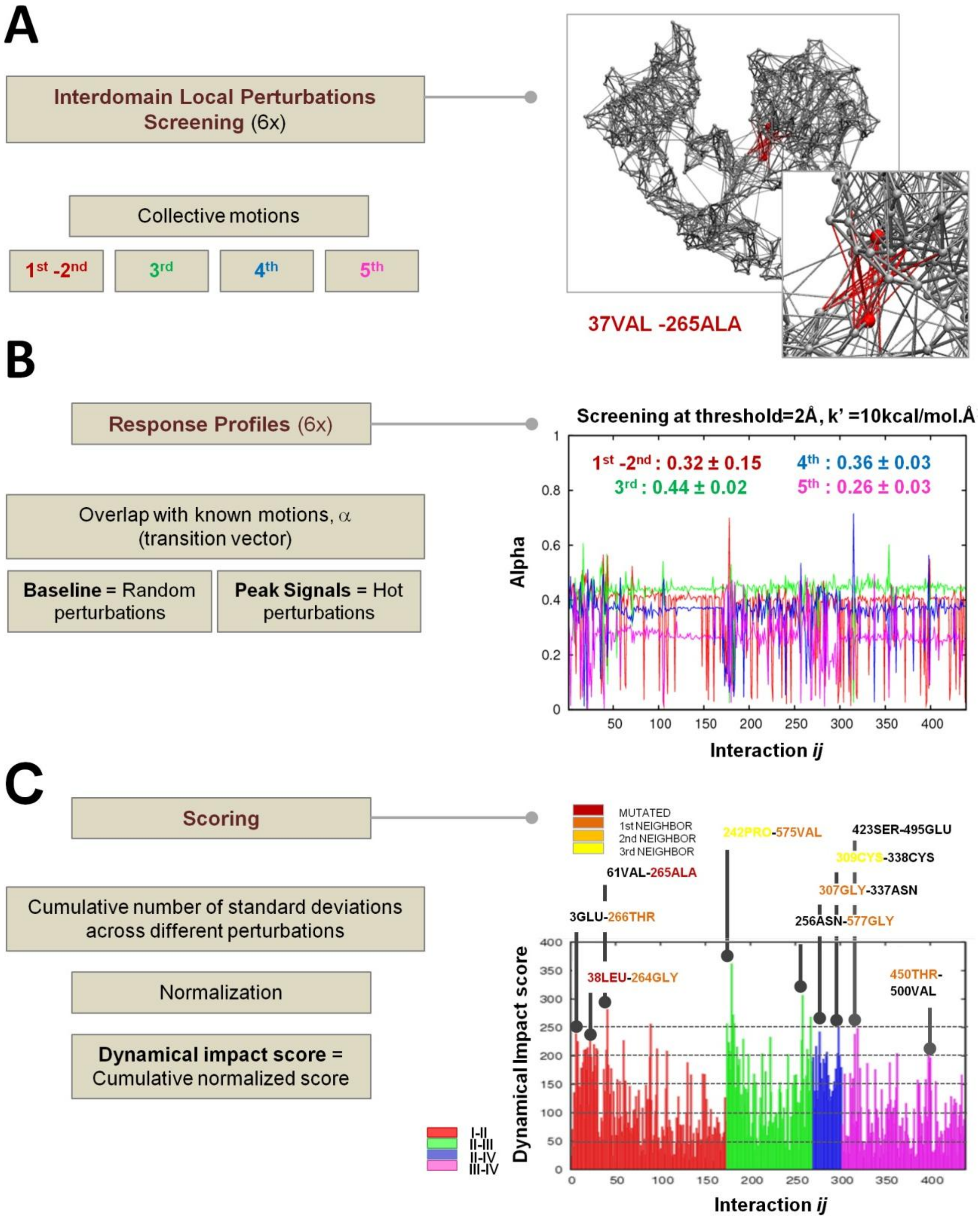


Figure 4. 2 **Elastic-Network perturbation screening of the interdomain interactions for closed sEGFR.** **A) Schematic representation of the perturbation method.** In red, residue pair defining the target spring and all its space neighbors for the pair 37VAL-265ALA at cutoff 4Å. The elastic network defined by ENM appears in gray; for the sake of clarity, only springs with values greater than 5kcal/mol.Å^2 are represented. **B) Response profiles based in the overlap (α) of the ENM collective motions with the known transition.** The responses of ENM-networks to perturbations across the interdomain interactions *ij* for a given screening is represented for each lowest-energy motion (*modes 1-2=red, 3=green, 4=blue, 5=purple*). Average overlap and standard deviation of each group of modes is indicated in the same colors. **C) Normalized Standard Deviations Sum** for the first 5 modes averaged over 6 different perturbative experiments (see *Extended Methods*). Most clustered peaks in the 1st-2nd top bins (scores over 200, *Table 4.3*) involve either mutated residues (red) or their nearest neighbors (orange-yellow scale), and target the same sites on 3D space.

Molecular Dynamics and Docking. We simulated both WT (unbound and bound) and mutant R84K sEGFR in closed and open states. Proteins were titrated, neutralized, hydrated, minimized, heated and equilibrated, and four replica trajectories for the six conditions were collected for 50 ns using the all-atom AMBER99SB-ildn force-field with GROMACS (Hess, Kutzner, Van Der Spoel, & Lindahl, 2008; Lindorff-Larsen et al., 2010), extending two of them up to 0,2µs (closed) or 0.1µs (open). Three of the closed state simulations were finally extended to 0,5µs, summing up a total simulated time of 2.5µs (See scheme and list of simulations in *Figure 4. 3* and *Table 4. 6*). Further simulations of other mutations were performed for the close state. Together, the total simulated time for sEGFR is near 5 µs distributed in more than 30 MD runs.

a) Intermediate sEGFR WT simulations: In order to validate that the untethered state can be stable for the WT sEGFR we took a representative untethered configuration reached by the R84K mutant and reverted it to the WT sequence. The system was then subjected to five replicas of 50 ns unrestrained MD simulation. Only in one case the protein re-tethers back to the closed state at the end of the simulation, confirming that untethered state is possible for the WT protein (even though less probable that for the R84K mutant) and that the tethered/untethered transition is reversible.

b) Intermediate sEGFR ligand binding in R84K and WT state: EGF was docked by *in house* tools to domain I of WT and R84K intermediate states of sEGFR using as binding mode that in the 1NQL structure, and was then subjected to five 50ns replicas. Ligand remained bound to domain I and also interacted simultaneously with domain III.

c) Intermediate sEGFR-Antibody complexes simulations: mAb806 and maAb175 antibody complexes with the sEGFR epitope were taken from PDB coordinates 3G5V and 3G5Y respectively and used to guide docking of the exposed epitope of our putative intermediate from the R84K simulation. The selected docked pose was then subjected to MD (5x 50ns trajectories), which confirmed that the intermediate conformation is stable and well suited for the recognition of the cryptic epitope by antibodies (impossible for known closed and open states).

d) Intermediate sEGFR dimer structures: We used as reference the active configuration found in *3njp* dimer to guide docking of the dimerization arms exposed in two intermediate structures, obtaining an unbound homodimer without steric clashes. The same docking/MD relaxation (5x 50ns replicas) procedure again revealed stable intermediate structures.

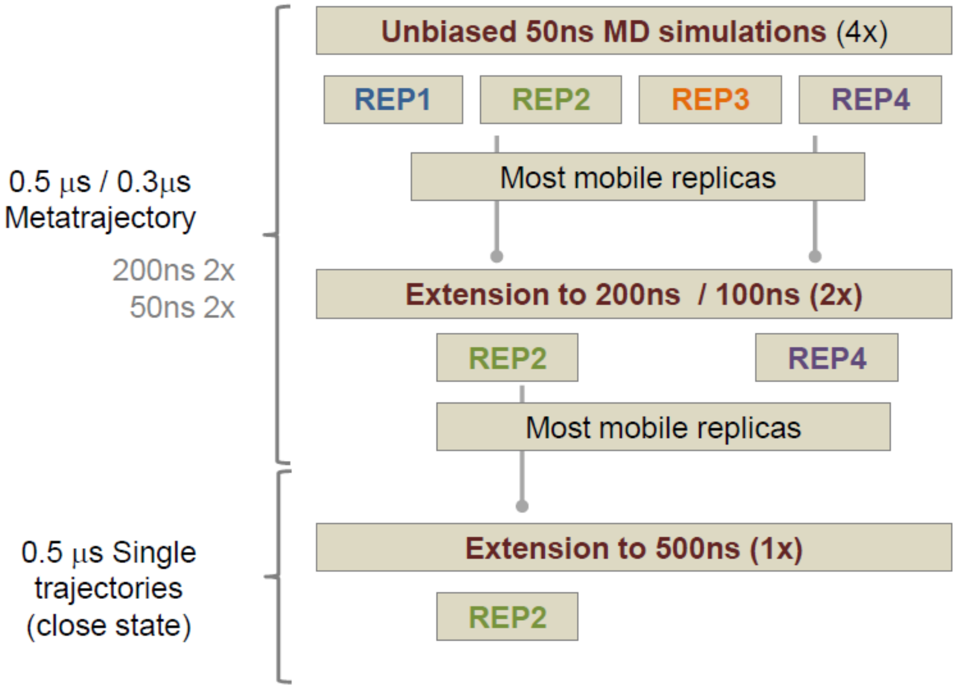

Unbiased 50ns MD simulations (4x)
REP1
REP2
REP3
REP4
Most mobile replicas
0.5 μs / 0.3μs Metatrajectory
200ns 2x
50ns 2x
Extension to 200ns / 100ns (2x)
REP2
REP4
Most mobile replicas
0.5 μs Single trajectories (close state)
Extension to 500ns (1x)
REP2

Figure 4. 3 **Enhanced sampling by tree-extension of MD trajectories to the submicrosecond range.** The scheme was repeated for WT and R84K mutant; in the last one, two replicas were extended to 500ns.

Molecular Dynamics Analysis.- The noise arising from short-range vibrations is filtered by *Essential Dynamics*, ED (Amadei et al., 1993) to obtain Principal Components (PCs) representing the essential movements. We use AMBER and GROMACS utilities to compute PCA, rMSD, rMSF, R_g and clusters. Hydrogen-bond analysis was performed using VMD defaults to detect bonds present in at least 40% of the trajectory. For analysis the noise arising from short-range vibrations is filtered by *Essential Dynamics*, ED (Amadei et al., 1993) to obtain Principal Components (PCs) representing the essential movements. To monitor structural changes, we use C^{α}-C^{α} distances and angles definitions for comparison with ENM generated structures.

a) Domain II curvature: we define it by the angle between C^{α} positions at the basis of the dimerization arm and the two extremes of domain II (residues 190-260-309)*;*

b) Domain I-III distance: in the unbound state, is monitored by the C^{α}-C^{α} distance for the salt bridge LYS13-ASP364, which is characteristic of ≈10^2 ns samplings for both R84K and WT sEGFR*;* in the bound state, is monitored by the C^{α}-C^{α} distance for the EGF-sEGFR salt bridge LYS48-ASP321, which approaches and stabilizes domains I-III surfaces through bound EGF

c) Domain II-IV tether distance: is monitored by the CA-CA distance between the key hydrogen bond TYR251-GLU578.

Note that other alternative angle and bond definitions are possible yielding similar results to track conformational changes.

Comparison metrics. Once obtained the principal motions of the structure (the set of modes, i.e. eigenvectors, ν_k, and eigenvalues, λ_k, from either ENM or MD, see details in **Chapter 3**), we measured the similarity and correlated motions:

1) Overlap of essential motions with the observed X-ray conformational change. The similarity between the essential deformation of the protein (in open or close state) and the transition is measured for each mode using a variation of Hess metrics (*Eq. 3.2*):

$$\alpha_k = \frac{\Delta r \cdot \nu_k}{\|\Delta r\| \|\nu_k\|}$$

(Equation 4. 4 *Overlap to transition vector*)

where $\Delta r = (R_2 - R_1) / \left\| R_2 - R_1 \right\|$ is the unitary transition vector between the two sets of coordinates, R_1 and R_2, describing the observed states of the protein (close and open) and ν_k is the *k-th* essential deformation mode. Generalization of *Eq.4.4* for the m-important deformation modes (the minimum set explaining a given threshold of variance) yields to a similarity index ranging from 0 (no similarity) to 1 (perfect similarity) (F Tama & Sanejouand, 2001; L.-W. Yang et al., 2009b):

$$\delta_k = \frac{1}{m} \left(\sum_{k=1}^{m} \alpha_k^2 \right)^{1/2}$$

(Equation 4. 5 *Squared Cumulative Overlap*)

2) Overlap between the essential motions computed from ENM and MD. Hess's metric (Hess, 2000) is used to estimate the similarity of ENM and MD-spaces for the first m-important modes

3) Cross-correlated movements. The covariance matrix describing the residue-residue correlations is computed from (Van Wynsberghe & Cui, 2006)

$$C_{ij} = \frac{\langle \Delta R_{\mathrm{i}} \cdot \Delta R_{\mathrm{j}} \rangle}{\langle (\Delta R_{\mathrm{i}})^2 \rangle^{1/2} \langle (\Delta R_{\mathrm{i}})^2 \rangle^{1/2}}$$

(Equation 4. 6 *Cross-correlation matrix*)

4.3 Unraveling dynamical hot spots: ENM analysis and network perturbation

4.3.1 Normal Mode Analysis of the HER1 ectodomain

The ectodomain is intrinsically prepared to perform close←→open transitions. The structural transition between close (1NQL (Ferguson et al., 2003)) and open (3NJP (C. Lu et al., 2010)) states of sEGFR is dramatic (rMSD=25.14Å), and involves a complex series of rigid domain movements, with minimal local changes in all domains except domain II, which bends as a spinal backbone to allow dimerization.

ENM analysis reveals that these motions are intrinsically imprinted in the ectodomain structure. The lowest-energy ED-ENM *normal modes* of the close state predict large relative rotations of the ligand-binding domains accompanied by bending of domain II (*Figure 4. 4B, top*), similar to those in known crystal structures. In fact, most of the *close → open* conformational change (75%) is explained by just the five lowest frequency modes (82% considering 10) (See *Table 4. 2*). The modes better aligned with the transition (the 1st, 3rd and 4th, with overlaps around 0.4) involve the inward rotation and apposition of the ligand binding domains, as well as tether extension (*Figure 4. 4A, double arrow*).

However, the reverse *open → close* conformational change maps even better with the collective modes from the open state (80% for 5 modes and 84% for 10, *Table 4. 2*). In contrast with the close form, which requires several modes to be described, now there is a major contribution of the 1st mode (overlap 0.75) guiding the closure of domains III-IV (*Figure 4. 5, A and B, top*). However, when EGF is introduced, the overlap between the open → close conformational change and the intrinsic motions of the open form significantly drops (grey line in *Figure 4. 5B, top*), suggesting that once the ligand is bound in its high-affinity configuration, may help to keep the receptor in the active state. Altogether, ENM suggests that sEGFR is intrinsically prepared to switch between close and open states, but the *close → open* transition needs the recruitment of several modes whereas the *open → close* inactivating change is easier and described by a single one, blocked by ligand binding.

Table 4. 2 Overlaps between the ED-ENM lowest frequency motions and the experimental transition vector for the 1nql ↔ 3njp conformational change in both directions.

PDB	overlap (5)	overlap (10)	Best	Rank	N_{evec} 90%
Closed (1nql)	0,75	0,82	0,44	3	8
Open (3njp)	**0,80**	0,84	0,58	1	11

Molecular joints control the transition between active and inactive states of the receptor. To better identify the key regions acting as hinges or "molecular joints" to transmit the collective motions, we analyzed the ENM-derived residue root mean square fluctuations (rMSF) and residue-residue correlations matrices (see *Methods*). Residue fluctuations show that in the close form, domains II and IV define a rigid region with coupled movements (*Figure 4. 4, left, red circle*) of the dimerization arm (*Figure 4. 4B bottom*, *green box*) and the *tether* (*same figure*, *blue box*); this region is surrounded by highly mobile modules, which allow domain II bending (*Figure 4. 4C*). Hinge-like interdomain points (seen as rigid sites next to high mobility segments in domain I; *Figure 4. 4B bottom*, *yellow box*) define a surface that acts as a "*molecular joint*" for the oscillations of domain I over domain II (*Figure 4. 4A,* yellow *star*).

This molecular joint target a series of small β-strands (Val36-Leu38, Glu60-Ala62, and Leu80-Arg84) at the basis of domain I, where glioma mutations appear clustered (*Figure 4. 4C*). The analysis of the residue-residue correlated motions contacts at this I-II interface transmit mechanical information from the mobile domain I to domain II, which in turns transfers motions to domain III (*Figure 4. 6, left,* note the block formed by domain I-IINtal (*green square*) and their sparser correlations with IICtal, coupled to domain III (*orange square*)).

Contrary to the complex pattern of natural motions in the close form, intrinsic mobility of the open state is simpler: the most flexible parts are the dimerization arm and the tether-Ctal portion of domain IV (*Figure 4. 5B bottom, blue* and *green boxes*); considering that the IV tail is fixed by the membrane, this movements may help to find the proper orientations for dimerization or retethering (*Figure 4. 5A, 1st mode*). Here the "*molecular joint*" between domains I and II appears again as a key region to control the coupled movements of both domains (*Figure 4. 6, right, green square*), which transfer the mechanical information to the rigid block formed by domains III-IV (*Figure 4. 6, right, orange square*). Altogether, these movements define an intrinsic pattern of motions that contains the biologically-relevant conformational changes, and suggest a key role for domain I-II interface in the control of the overall dynamics. A second aggregation of hinge points defines another "*molecular joint*" in domain III, around Glu388-His394, 400Glu and Leu419-Leu429 (see *Figure 4. 4B bottom*), which controls minor displacements with respect to domain IV (*Figure 4. 4)*, and also contains a few reported mutations (see *below*).

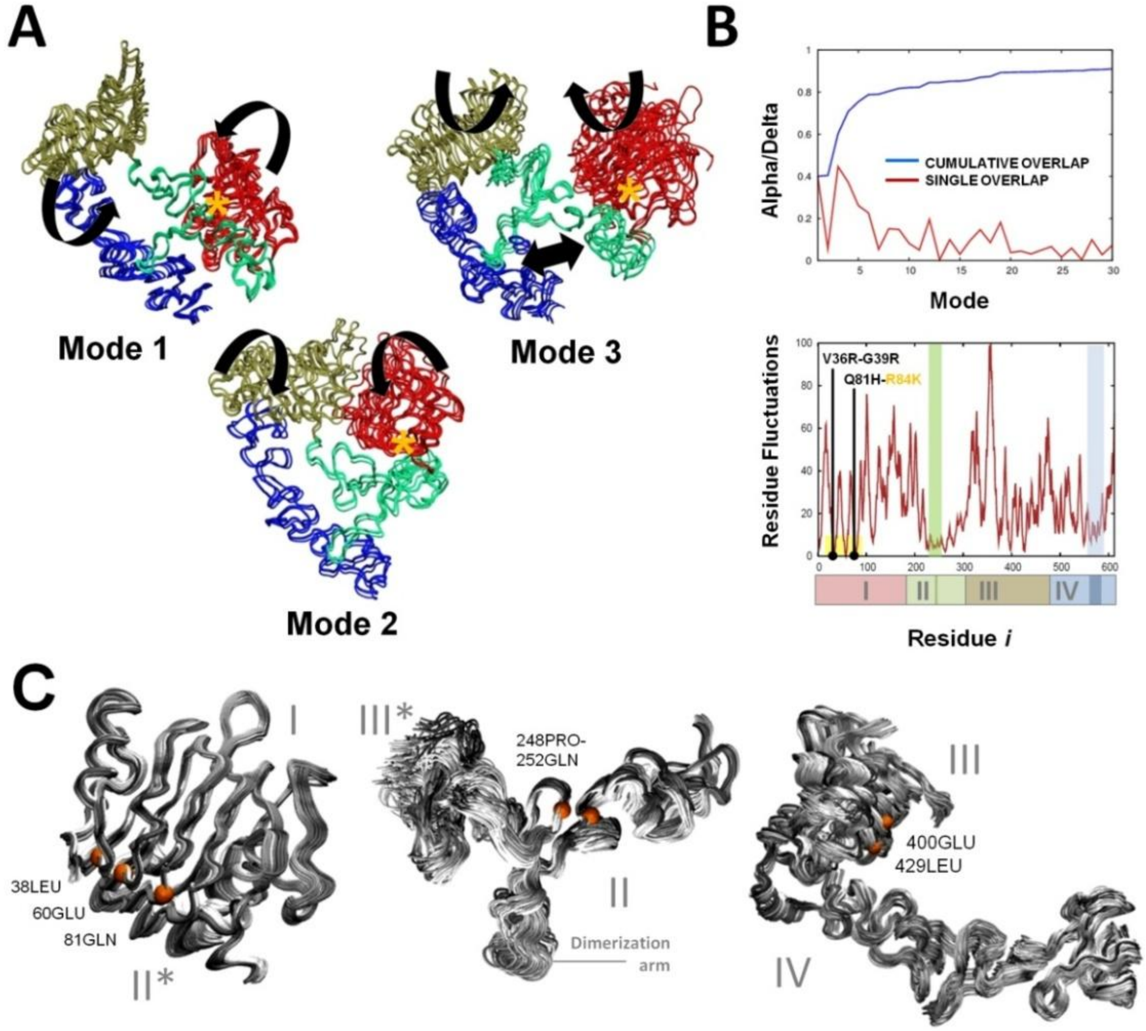


Figure 4. 4 **The Elastic Network Model of closed sEGFR encodes the transition towards the open state.**

A**) Structural ensembles along the first three modes of the sEGFR close conformation show large domain I-III rotations, domain II bending and untethering.** The directions of motion of the ligand binding domains I and III are indicated with curved arrows. Note the large domain I oscillations, which are coupled to I-III relative rotation movements, domain II bending and tether-breaking oscillations in mode 3 (*double arrow*). Domains I-IV colored *red-green-brown-blue*. **B) Motion pattern from ED-ENM.** *TOP:* Single (*red*) and cumulative (*blue*) overlap of the principal motions with the close-> open transition. *BOTTOM:* Residue Fluctuations for the first 10 modes; rigid sites appear as minima and flexible regions as peaks. The dimerization arm/tether are marked with green/blue boxes to facilitate discussion and some oncogenic mutation clusters in the domain I-II hinge surface (*highlighted in yellow*) are shown as solid black circles **C) A focus on the hinge-like points acting in the collective motions of closed sEGFR.** A structural ensemble for the first 10 modes is represented, with residues acting as hinges highlighted as orange spheres. *Left:* residues at the hinge surface in the base of domain I, contacting with domain II (*II*, not shown*), contain the 3X- and 8X- mutation clusters (see main text). *Middle:* domain II is also bent around the basis of the dimerization arm; note that bending motions are focused in the loop connected to domain III (*III*, not shown*). Right: domain III-IV interdomain surface showing hinge points, which also target some mutations (see main text).

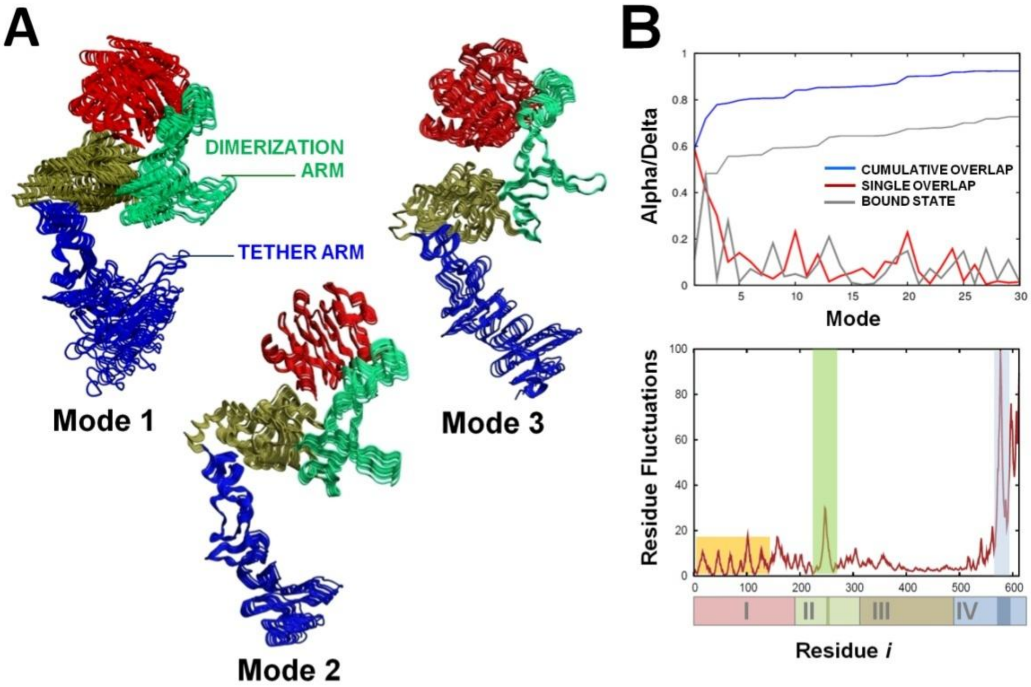


Figure 4. 5 **The Elastic Network MOdel of the sEGFR open state describes the inactivating transition.**

A) Structural ensemble along the lowest-frequency modes of the sEGFR open conformation. Note the large oscillations in the first mode approaching domain II (green) dimerization arm and domain IV (blue) C-terminal tail (including the tether arm). **B) Elastic Network Normal Mode Analysis of the flexibility of sEGFR open state.** *TOP:* Single (red) and Cumulative (blue) overlap of essential movements with the open→ close transition. Note how similarity drops when ligand is introduced in the elastic network (gray plot), indicating a more difficult inactivating open→close transition. *BOTTOM:* Residue Fluctuations (first 10 modes). Note that the dimerization arm (*green box*) and the C-terminal part of domain IV including the tether (*blue box*) are very flexible.

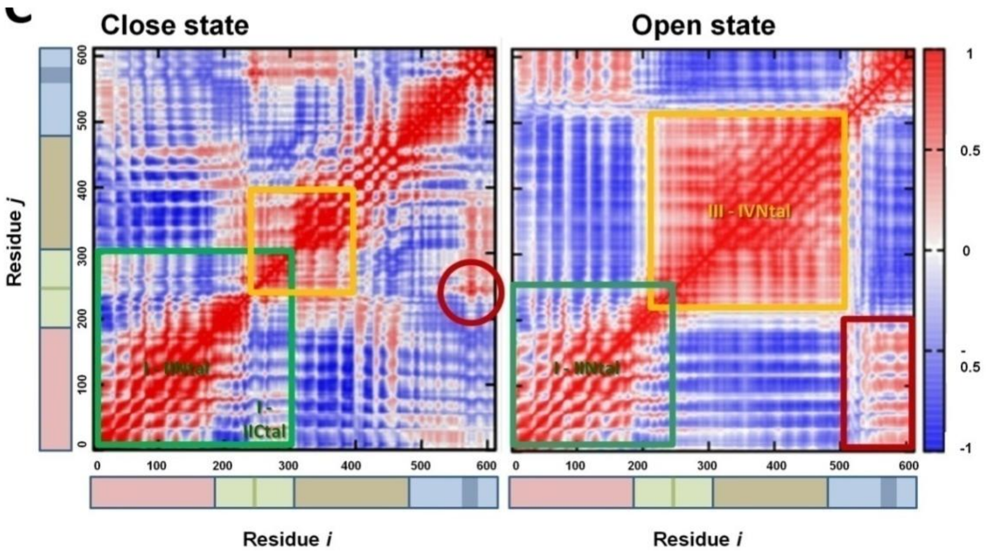


Figure 4. 6 **Residue-residue cross-correlations for the closed (*left*) and open (*right*) states.**

Positive correlations appear in red and negative ones in blue. *LEFT:* In the closed state, mechanical oscillations of the domain I can be transferred to domain II and the dimerization arm (green box) and through domain II to domain III (*yellow square*) and the tether arm (*red circle*). *RIGHT:* In the open state, the extent of domain I-II correlations (*green square*) is smaller, and domain I moves coupled mostly to domain II Ntal (labeled I-IINtal) and domain IV Ctal (*red square*); the rest of the protein forms a large block (domains IICtal-III-IVNtal, *yellow square*).

4.3.2 Local perturbation methods: hot spots target mutation clusters

Perturbation Analysis identifies dynamical hot spots targeting interdomain mutations. The ENM results suggest that mutations, especially at hinge interdomain surfaces, could interfere with the close ectodomain dynamics. We applied our ENM-perturbation scheme (see *Methods* and *Figure 4. 2*) to detect structural regions whose perturbation can have a strong effect in the biologically-relevant intrinsic pattern of motions. We screen all interdomain interactions by perturbing one by one each spring plus its local neighbors and measure the effect in the similarity between the dominant lowest frequency modes and the vector describing the transition. Dynamically critical regions are detected by computing the average standard deviation across different perturbative experiments for each screened interaction. The screened interactions are colored by the number of standard deviations versus the overall average in *Figure 4. 7.* As can be observed, critical interactions appear clustered in three-dimensional space and display a high correlation with hot mutation sites, which are also highly conserved.

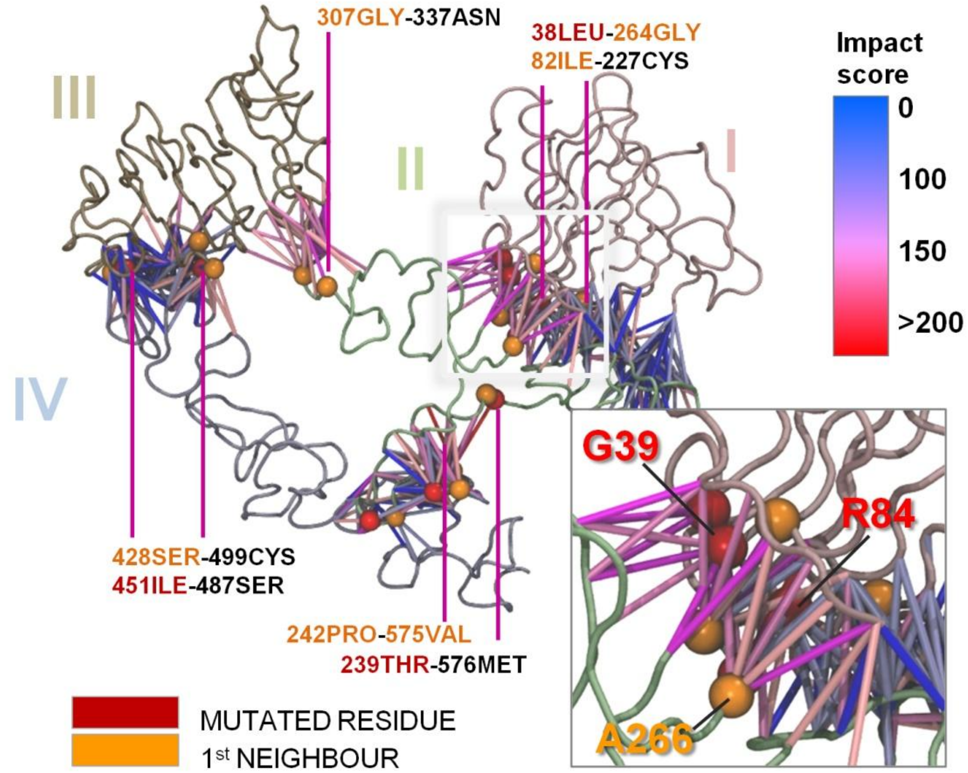


Figure 4. 7 **Spatial clustering of the dynamically hot interactions targets mutated sites.**

Springs colored by dynamical impact score for the first m=5 modes (ζ_{ij}^{m}, see *Methods*); mutated residues or their nearest-neighbors connected by high-scoring springs are represented as red and orange balls, respectively. Note the clustering of top-scoring springs (*red-magenta*) in regions that concentrate oncogenic mutations. A close-up to the hot spot at interdomain I-II hinge interface showing clustered *glioblastoma* mutations is shown in the lower right corner. Domains I-IV colored pale pink-green-brown-blue.

Most of the sensitive contacts detected by the analysis (*Table 4. 3*) appear clustered (red-pink pseudobonds in *Figure 4. 7*) at well-defined sites coincident with cancer-mutation spots. As could be anticipated, the interface of domains II-IV (*tether*) - which concentrates known mutations whose functional impact is clear (such as G574V or T239P) - is a perturbation-sensitive region (highest peak in *Figure 4. 2C*). Other hot region is the Cys-loop connecting the II-III interface, which targets a large mutation cluster in HER1 (G298C-E306K) and its *C.elegans* homolog LET-23 (Katz et al., 1996), and plays a key dynamical role in HER4 (Du et al., 2012). At the domain I-II hinge surface, the hottest spots involve interactions between the Val36-Gly39 (V36L-L38R-G39R cluster) and the most prevalent glioma mutation, Ala265 (A265V/D/T), followed by neighbor contacts between the Gln81-Arg84 (Q81H-R84K cluster) and the opposite domain II surface (*Figure 4. 7.,* detail) which mediate key domain I-II dynamical coupling. Further hot spots appear in the hinge surface between domains III-IV including non-*glioma* mutations such as R427F and I451V (in lung and head/neck carcinomas). This suggests that mutations at interdomain hinge-like regions and loops perturb the intrinsic functional motions, and outlines a new oncogenic mechanism based on the alteration of protein dynamics.

Table 4. 3 Top scoring peaks at the Local Perturbation Screening considering the first five modes. In column 1, mutated residues are highlighted in bold letters. nSD sum = nSD across the perturbation experiments added for the first 5 modes; NnSD sum= normalized ASD across the perturbation experiments added for the first 5 modes.

Pseudo-bond		Peak Group	nSD sum	NnSD sum	Sequence-related mutations
242PRO	577GLY	**II-IV (tether)**	9.2	319	T239P
254ASP	**578GLU**		6.9	206	57X-cluster
243LEU	576MET		6.8	188	
241PRO	577GLY		6.8	210	
256ASN	577GLY		6.7	142	
257PRO	579ASN		6.2	195	
245LEU	**578GLU**		6.0	160	
244MET	576MET		5.7	200	
309CYS	337ASN	**II-III hinge**	8.6	238	30X-cluster
308PRO	375LYS		7.9	197	
307GLY	337ASN		5.9	169	
308PRO	310ARG		5.7	218	
447TYR	483HIS	**III-IV hinge surface**	8.3	241	R427F
424LEU	492TRP		7.2	155	I451V
451ILE	492TRP		7.2	195	
427ARG	497ARG		6.8	163	
424LEU	491CYS		6.2	164	
427ARG	495GLU		5.9	158	
481VAL	482CYS		5.1	172	
4LYS	266THR	**I-II hinge surface**	7.8	226	3X-cluster
37VAL	**265ALA**		7.2	177	8X-cluster *
38LEU	266THR		7.0	162	26X-cluster *
187THR	207CYS		6.7	192	
187THR	208CYS		6.2	166	
115ASN	214ALA		6.1	140	
84ARG	213ALA		5.7	134	
37VAL	264GLY		5.7	130	

4.4 Molecular Dynamics in the submicrosecond scale

Molecular dynamics (MD) simulations reveal dramatic motions of the ligand-binding domains in closed sEGFR. To sample more accurately the flexibility of HER1 we performed multiple unbiased MD simulations of the unbound close state and as control, also in the open one, extending some of them to the near-microsecond timescale (see *Methods*). Trajectories show that both conformations are quite flexible, displaying a pattern of motions that fits perfectly with ENM modes and with the open⇔close change (overlaps of 81% (close) and 87% (open) between the first five modes and the transition) (see *Table 4. 4*).

Table 4. 4 rMSD and Essential Dynamics of Metatrajectories.

Overlaps between MD Principal Components and the transition vector for the 1nql ↔ 3njp conformational change (in brackets: similarity to ED-ENM low frequency motions). 500ns-long trajectories not included.

STATE	Total Time	Variant	Avg rMSD	rMSD Max/Min	Avg Rg	Rg Max/Min	α_5	α_{10} (α_{NMA})	α_{max}(α_{NMA})	rank	Number of Clusters
***Closed* (1nql)**	**500ns**	*WT*	8.22 ± 1.74	14.00 / 1.20	35.29 ± 0.76	38.84 / 32.50	0.81	0.85 (0.69)	0.60 (0.56)	2	9
		WT+EGF	6.62 ± 1.75	14.25 / 1.06	34.20 ± 1.12	38.14 / 27.92	0.86	0.91 (0.77)	0.78 (0.84)	2	10
		R84K	8.68 ± 3.02	14.93 / 1.34	32.57 ± 2.46	38.64 / 30.95	0.77	0.91 (0.72)	0.46 (0.58)	3	10
***Open* (3njp)**	**300ns**	*WT*	5.29 ± 1.24	11.03 / 1.29	33.92 ± 0.60	36.68 / 31.35	0.87	0.89 (0.84)	0.69 (0.74)	3	3
		WT+EGF	2.79 ± 0.71	6.54 / 0.98	34.83 ± 0.47	36.70 / 33.05	0.83	0.84 (0.72)	0.69 (0.92)	2	1
		R84K	4.44 ± 0.83	8.30 / 1.15	33.95 ± 0.62	37.02 / 32.07	0.85	0.86 (0.83)	0.82 (0.54)	2	4

MD simulations also confirm that the closed WT sEGFR is extremely flexible and samples spontaneously a wide range of conformations (*Figure 4. 8,* note the rMSD spanning 8Å). This surprising mobility is mainly due to large oscillation motions of domain I over domain II (*Figure 4. 8 top*, head movements marked with arrows), where the interface between both acts as a hinge as predicted from ENM (*Figure 4. 9A top*, highlighted in yellow). The domain I oscillation motions lead in some sampled structures to dramatic increases in domain I-III distances, stretching domain II and loosening the tether-dimerization arm contacts (see for example *Figure 4. 10B*, *green cluster 6*). All these movements are enhanced as trajectories extend to the microsecond range *Figure 4. 12A*), with the domain I-III cleft increasing up to 60Å (see *Figure 4. 13*A, arrow in the 3rd column), domain II stretching to near 180 °, and the tether breaking transiently (see *Figure 4. 13*A, 2nd and 3rd columns). After each stretching event, domain II relaxes and bends again, finally triggering a collapse of the ligand-binding domains at the end of the simulation (*Figure 4. 12A,* last frame). This remarkable stretching/bending mechanics of domain II is allowed by the plasticity of the hot Cys-loop which connects it to domain III, as noted in HER4 by (Du et al., 2012).

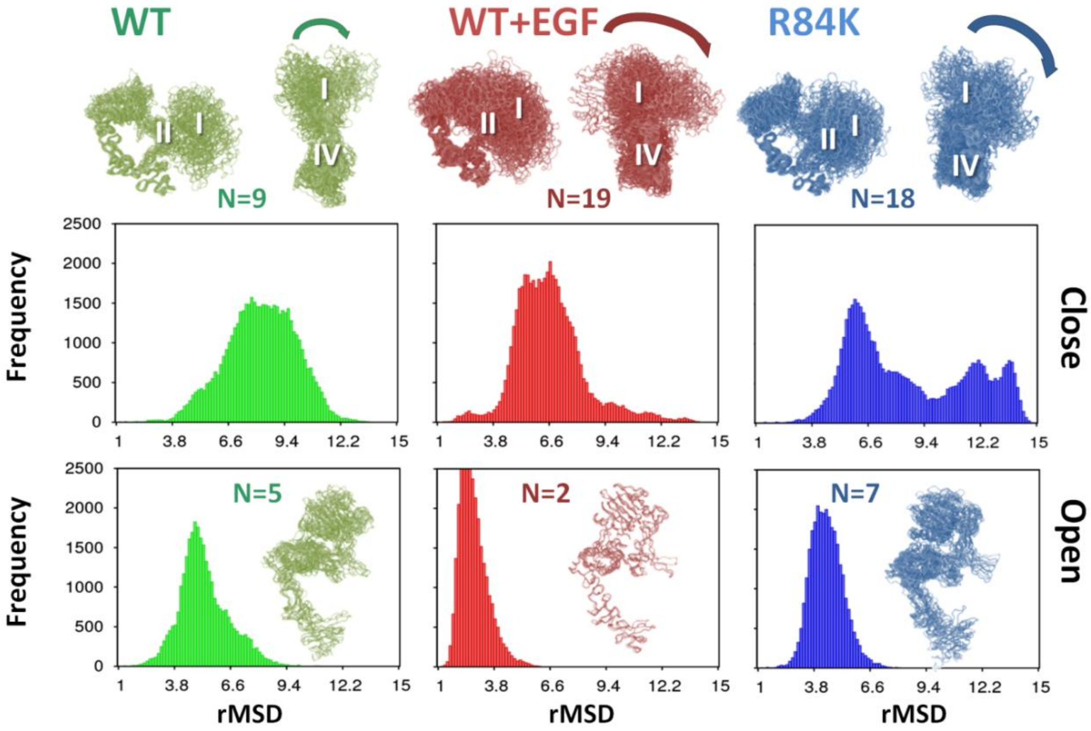


Figure 4. 8 **Root mean square deviations (rMSD) distribution for MD metatrajectories sampling 0.5μs for the closed state (*upper row*) and 0.3μs for the open state (*lower row*), with respect to the starting crystal structures.**

Aligned clusters from metatrajectories appear in *green* (Wild-Type, WT, unbound), *red* (WT+EGF, bound) and *blue* (R84K mutant, unbound); for the closed state (*upper row*) a view from the front (*left*) and the bottom (*right*) is shown, to highlight domain I oscillation movements (*curved arrows*). Compare the amplitude of domain I oscillations in the bound and R84K states with the WT reduced motions. The mobility of the open state is remarkably lower than that of the close state, with a narrower rMSD distribution shifted to lower values, and a smaller number of clusters, especially in the bound state. *N*= average number of clusters for each state.

The observed relative large-scale fluctuation motions of the unbound ligand-binding domains agree well with tryptophan fluorescence data (Kozer et al., 2011) suggesting that domain I and III rotate freely until ligand binds, whereas tether stretching and loosening is also consistent with the low barrier for untethering suggested previously (Arkhipov et al., 2013; Du et al., 2012; Ferguson, 2009). However, in the absence of ligand these spontaneous movements never lead to an open-state, and the tether holds the dimerization arm hidden in the observed timescale, as in microsecond-long simulations (Arkhipov et al., 2013). As anticipated by ENM, the open structure displays less flexibility than the close form (note the narrower rMSD distributions in *Figure 4. 8 lower row*), mainly limited to oscillations of the free dimerization arm and the tether-domain IV Ctal region (*Figure 4. 9A bottom, green and blue boxes*), movements previously observed in long timescale simulations (Arkhipov et al., 2013). Domain I-II motions (*Figure 4. 9A bottom, yellow box*) however are smaller, as consequence of the greater stiffness and stability of this configuration even in the absence of ligand.

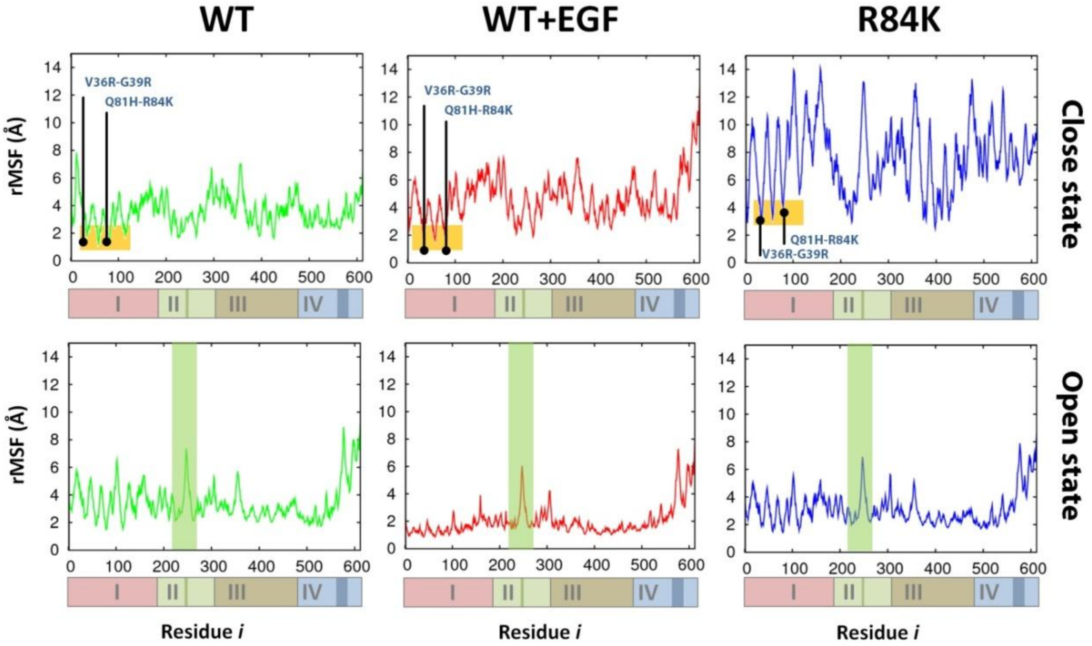


Figure 4. 9 **Residue root Mean Square Fluctuations (rMSF) in metatrajectories sampling 500ns for the closed state (*upper row*) and 300ns for the open state (*lower row*).**

For the closed state, domain I-II hinge surfaces are highlighted in yellow with mutation clusters as black dots. Note the overall increase in the flexibility of R84K mutant compared with WT (*blue profile*), especially at the I-II hinge surface and the tether (marked with a *green box*). See also the reduced flexibility of the close state in all cases (especially bound to the EGF).

Ligand binding drives the close state to an open but inactive conformation, while keeping rigid the active one. In order to characterize ligand-driven untethering, we performed several simulations of the ectodomain bound to domain I as appears in the crystal. EGF binding to the WT close state leads to an even better alignment of the intrinsic dynamics along the *close* → *open* transition pathway (*Table 4. 4*). In our trajectories, the ligand remains bound to the binding site at domain I, shifting its intrinsic oscillation motions over domain II to rotate and orthogonally approach domain III (*Figure 4. 10B, red cluster 2*). As domains I-III get closer, domain II is forced to bend, contrary to the free stretching movement predominant in the unbound forms of WT (discussed *before*) and R84K sEGFR (see *below*). After each bending event, relaxation of domain II separates again domains I-III. This alternative binding/unbinding of EGF to domain III coupled to bending/relaxation of domain II, is followed by loosening of the tether, up to the point that transient untethering is detected even in short trajectories (*Figure 4. 10B red cluster 13*, tether detail below).

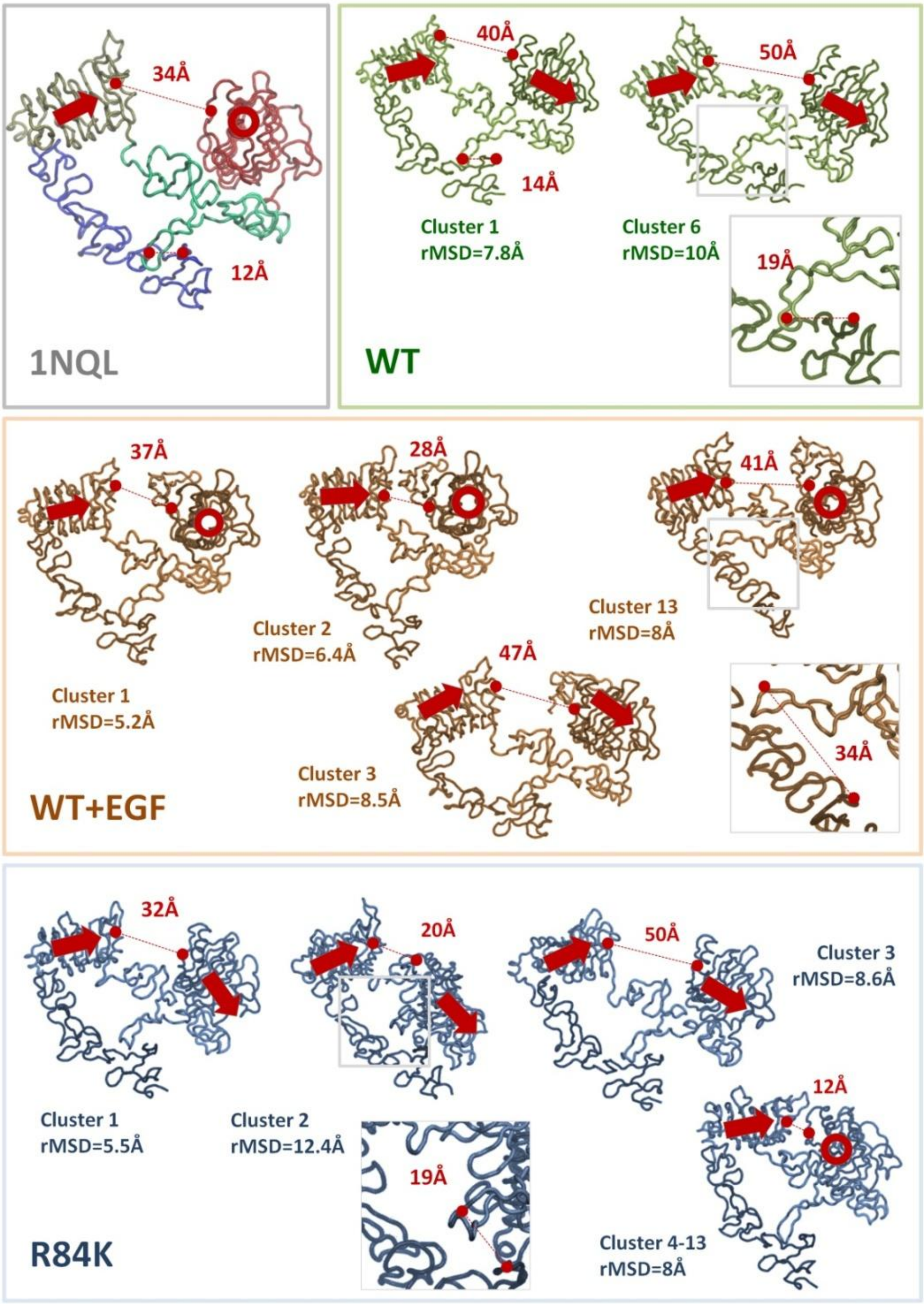


Figure 4. 10. **Major clusters and spontaneous untethering in closed state 0.5 μs metatrajectories.** *WT=Wild type sEGFR (green structures); WT+EGF=sEGFR with EGF (ochre structures); R84K mutant =sEGFR (blue structures).* The starting structure (1NQL) is shown as reference with domains I-IV red-green-ochre-blue. Interdomain Cα-Cα distances between domains I-III and domains II-IV appear in red (*dashed lines*); see *Methods* for definitions. The orientation of domains I and III are highlighted with arrows (N to C-tal); a circle indicates an arrow perpendicular to the page). Mutant and bound states are very polymorphic, with twice the number of clusters of WT sEGFR. Ligand-independent untethering is observed after domain I-III extension in WT state (*cluster 6*, detail) as well as in R84K (*cluster 3*), but in the last proceeds fast to an intermediate domain I-tilted state (*cluster 2*, detail) or to a collapsed I-III state (*cluster 4*). On the contrary, ligand binding approaches and orientates domain I orthogonal to domain III instead of parallel (*clusters 1-2*).

Analysis of a 0.5 μs simulation (*Figure 4. 12B*) clearly shows how the ligand triggers periodic bending of domain II (see *Figure 4. 13B, 2nd column grey arrows*), coupled to successive and longer tether opening events. When irreversible untethering finally happens (*Figure 4. 12B*, 4th frame, *yellow circle* and also *Figure 4. 13B*), domain IV quickly departs from domain III. A was observed in HER4 (Du et al., 2012), such transition leads to an extended conformation similar to the open state (*Figure 4. 12B*, last frame), but that still needs a major rotation of domain IV to form the high-affinity binding site of the fully active conformation in the crystal dimer as we will discuss below. Finally, as predicted from ENM, MD suggests that EGF high-affinity binding in the open state next to the domain I-II hinge surface, blocks domain I motions (*Figure 4. 9A bottom, yellow box*) and stabilizes this conformation (note the extremely narrow rMSD in *Figure 4. 8 bottom*) reducing the possibility of the inactivating *open → close* transition. This stabilizing role of the ligand is in agreement with several experimental observations.

Table 4. 5 **Summary of persistent I-II interdomain hydrogen bonds during 500ns MD meta-trajectories**.
Note the disappearance of the key Arg84-Cys227 hydrogen bond in R84K. Results are reported as time percentage of the trajectory when bonds are present.

Domain I-II H-bond		*WT*	*R84K*	*WT+EGF*
ARG198-side	GLU118-side	95%	100%	90%
ARG84-side	CYS227-main	45%	0%	40%
CYS199-main	THRE187-main	31%	30%	30%
ARG231-side	NGL3-side	30%	18%	30%
ALA214-main	GLU118-side	22%	31%	60%
ARG198-side	THRE187-side	24%	23%	33%
THRE187-side	ALA214-main	28%	18%	38%
TYR275-side	GLN8-side	18%	11%	22%
ARG198-side	ASP142-main	20%	27%	11%
ARG228-side	GLU118-side	15%	8%	0%

MD simulations of the R84K mutant reveal an alternative ligand-independent transition leading to a closed untethered state capable to dimerize. ENM calculations suggest that oncogenic mutations affect the intrinsic motions of the close state, and thus, it is tempting to argue that they might favor spontaneous transitions to partially or fully untethered structures. To test this hypothesis we performed multiple MD simulations of the aggressive R84K mutation of sEFGR (which targets a domain I-II hotspot) in the unbound close state. Despite the apparently conservative nature of the change, the R84K ectodomain displays important changes in the close state dynamics compared to the WT protein. In all replicas a crucial hydrogen bond connecting the highly conserved Arg84 and Cys225 is lost (*Table 4. 5*) leading to a dramatic increase in flexibility, with multiple subpopulations in rMSD spanning 14Å (*Figure 4. 8top*).

This augmented flexibility of the mutant is due to a extremely floppy I-II interface, which allows the mutant to display *enhanced* WT-like motions: there are large-scale oscillations of domain I over domain II (*Figure 4. 9A top* and *Figure 4. 10top)* leading to extreme domain I-III rearrangements and variations in the dimension of the binding cleft (note changing I-III distances and orientations, *Figure 4. 10B*), accompanied by stable untethering (*Figure 4. 10,* blue clusters detail). Analysis of a 0.5 μs trajectory of the R84K mutant (*Figure 4. 12C*) confirms that spontaneous untethering is caused by maximization of the WT motions. The enlargement of domain I-III distance (near 60Å, see *Figure 4. 13*C) stretches domain II (up to 180^{o}, *Figure 4. 12C* and *Figure 4. 13)* but in the mutant, the floppy domain I rotates more freely and tilts back up to contact domain IV breaking the tether (see *Figure 4. 12C*, yellow circle and *Figure 4. 13C, star*). This sequence of movements, never observed in WT runs, generates the complete exposure of the dimerization arm in a closed but untethered conformation (*Figure 4. 16A*) which remains stable throughout the remaining 500ns simulation (*Figure 4. 13C)* and is remarkably similar to the first ENM mode which separates WT and mutant samplings (see *Figure 4.11* and *Figure 4. 14*). The most salient features of the intermediate identified are the evident exposure of both the dimerization arm and the mAb806/175 conformational epitope, as we will discuss later. We have detected virtually identical transitions in independent short MD runs of nearby mutations such as G39R, where the introduction of a charged residue disrupts a hydrophobic patch in the I-II interface (*not shown*). In both cases the intermediate appears in the first 50ns and remains stable during the rest of the simulation. Contrary to the general view, this mutant untethered and unbound state, different from all solved structures, is more similar to the close than to the open conformation. We also tested if this novel state is stable for WT sEGFR running five 50ns replicas of this configuration with the original sequence; only in one case the structure re-tethers back to the starting state, confirming that the tethered/untethered transition is stable but also reversible.

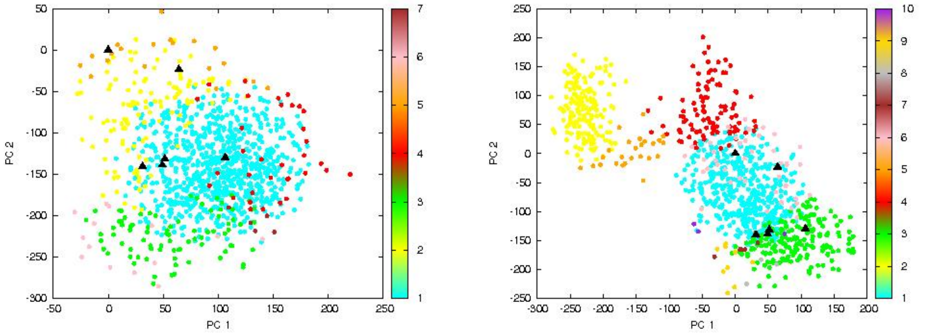


Figure 4. 11 **Projections of the major clusters for WT (left) and R84K (right) 0.5 μs metatrajectories onto the first two normal modes.**

The projection onto the normal modes clearly shows that mutant R84K sEGFR sample extensively the first principal component and populate a different region of the conformational space (cluster 2, in yellow, corresponds to the closed untethered state).

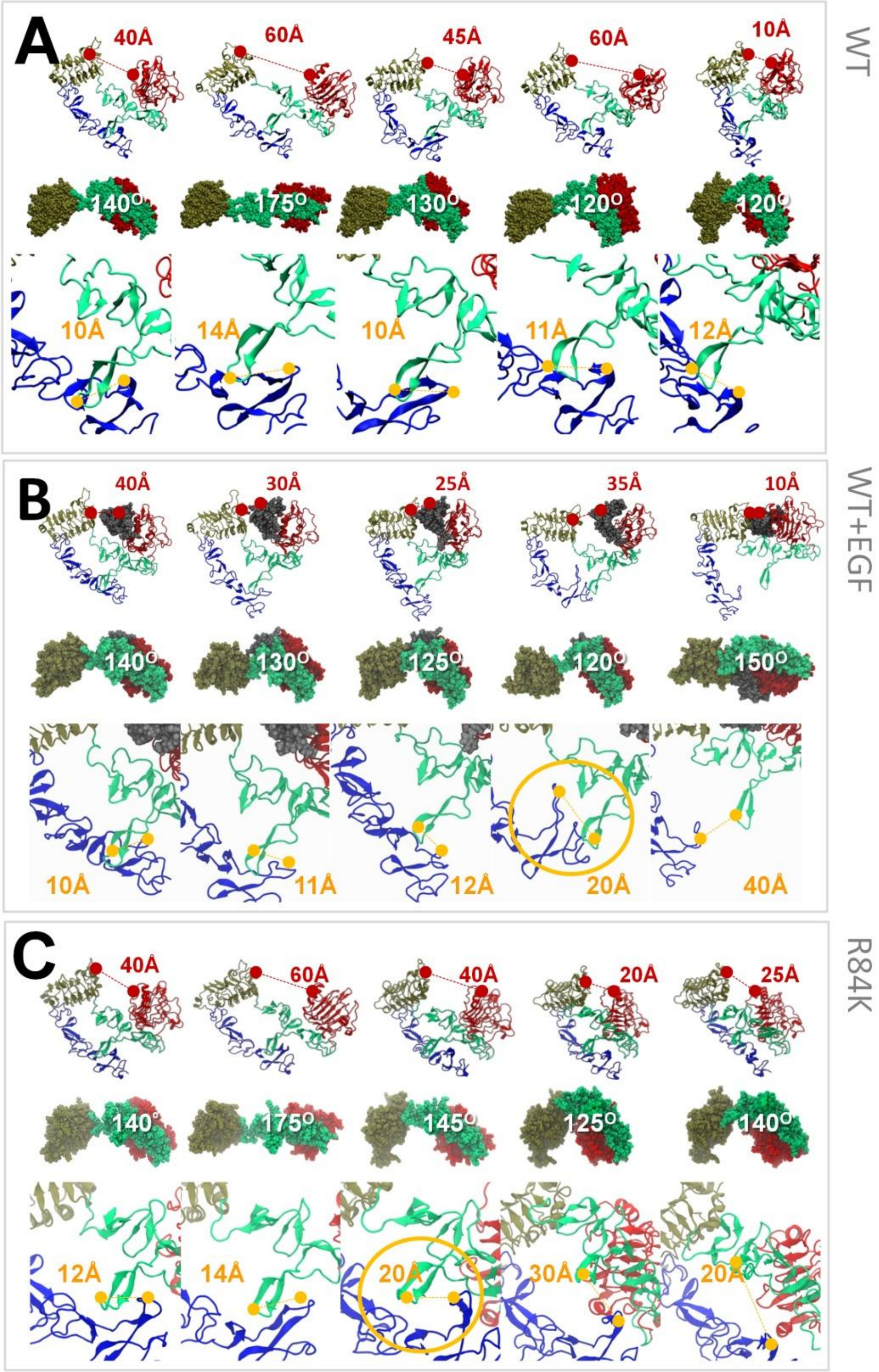


Figure 4. 12 **Representative structures obtained in very long MD trajectories display untethering transitions for unbound R84K (C) and bound WT sEGFR (B), but not WT sEGFR (A).**

Domains I-IV colored red-green-brown-blue. The starting state is the same for all simulations (*1st frame*). For each box A-C: Front view of representative snapshots with domain I-III distances in red (*upper row*); a view from below to highlight domain II bending coupled to changing domain I-III distances (domain IV removed) with the bending angle in white (*middle row*); close-up to the tether region, II-IV distance in orange with irreversible untethering events highlighted with a circle (*lower row*). **(A)** WT sEGFR undergoes large domain II stretching motions (~180°) coupled to domain I-III separation, followed by slow domain II relaxation and bending. **(B)** Ligand-driven untethering follows domain I-III apposition and repeated domain II bending. **(C)** Spontaneous untethering in the R84K mutant follows domain I-III separation and domain II stretching as in (A), but here the extreme flexibility of the I-II interface traps domain I in a tilted-back configuration breaking the tether. In both B-C, the tether is broken when domain II returns to its native state bending angle (~140-150°, *1st frames*).

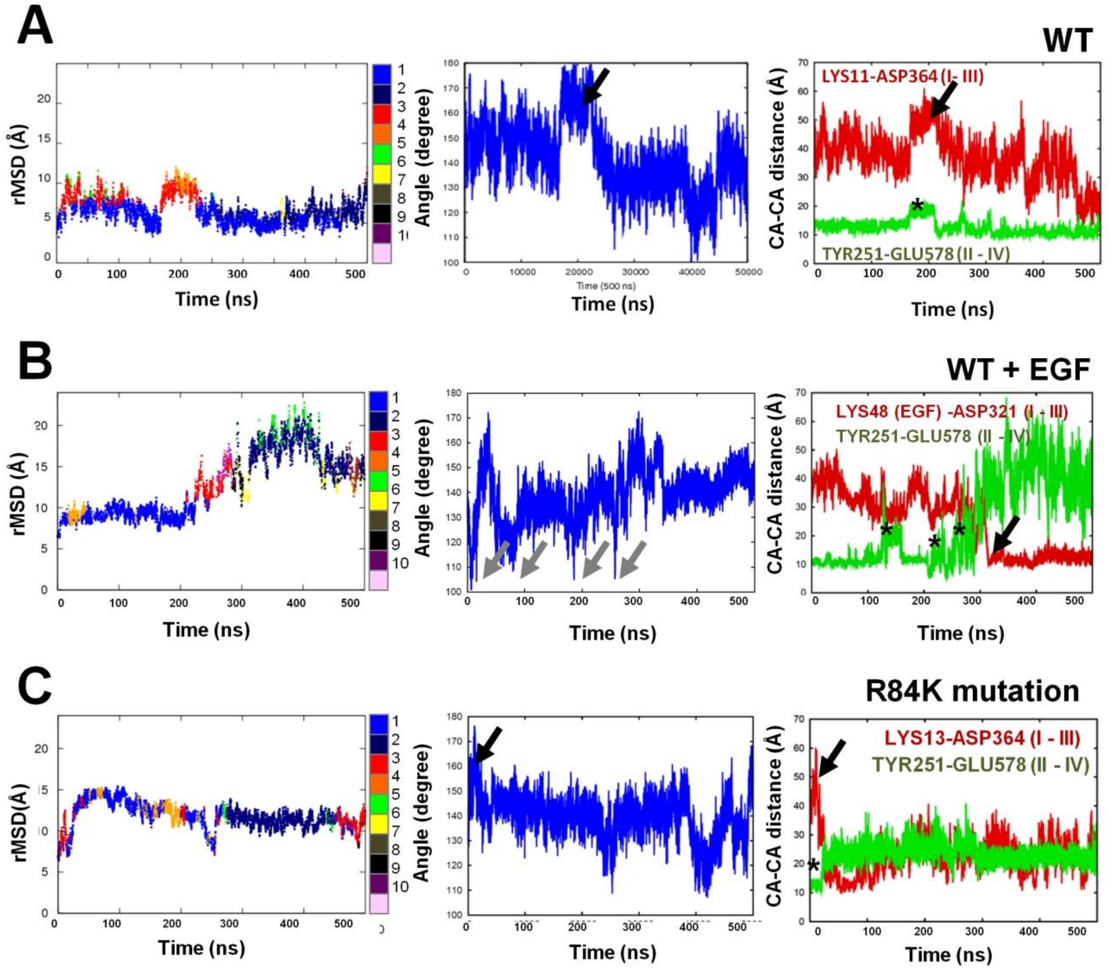


Figure 4. 13 **Extended 0.5 μs trajectories for WT unbound sEGFR (A), WT bound sEGFR (B) and R84K unbound sEGFR (C).**

From the first to the third column, geometrical descriptors are: rMSD versus initial state colored according to major clusters (1-10), domain II angle bending, and domain I-III and II-IV distances (see *Methods* for definitions). Untethering events are marked with stars in column 3 (*green plot*). Note how in WT (A) and R84K (C) domain II stretching peaks (black arrows in column 2) are associated with domain I-III separation (arrows in column 3, *red*) and breaking of the tether. Ligand binding (see B) forces domain II bending rather than extension (grey arrows in column 2), which is followed by increasing tether breaking events after periodic relaxation times (*column 3*).

4.5 Discussion: A novel mechanism for HER1 activation

As we mentioned above, the dynamics and oncogenic mutations of the extracellular domain have still not been analyzed in-depth. Therefore, in spite of the recent efforts to simulate the ectodomain (Arkhipov et al., 2013; Du et al., 2012), key aspects of the activation process at the extracellular level are still poorly understood. It is widely assumed that, either triggered by ligand binding or spontaneous, untethering and receptor opening are parallel *i.e.* that untethered states have an extended configuration similar to that found in crystals for bound HER1 or unbound HER2. Therefore, activating mutations are expected to shift the ectodomain towards open untethered states. Here, the combination of coarse-grained and atomistic simulations has revealed, on the contrary, that ***mutations shift the dynamics towards an untethered but closed state***.

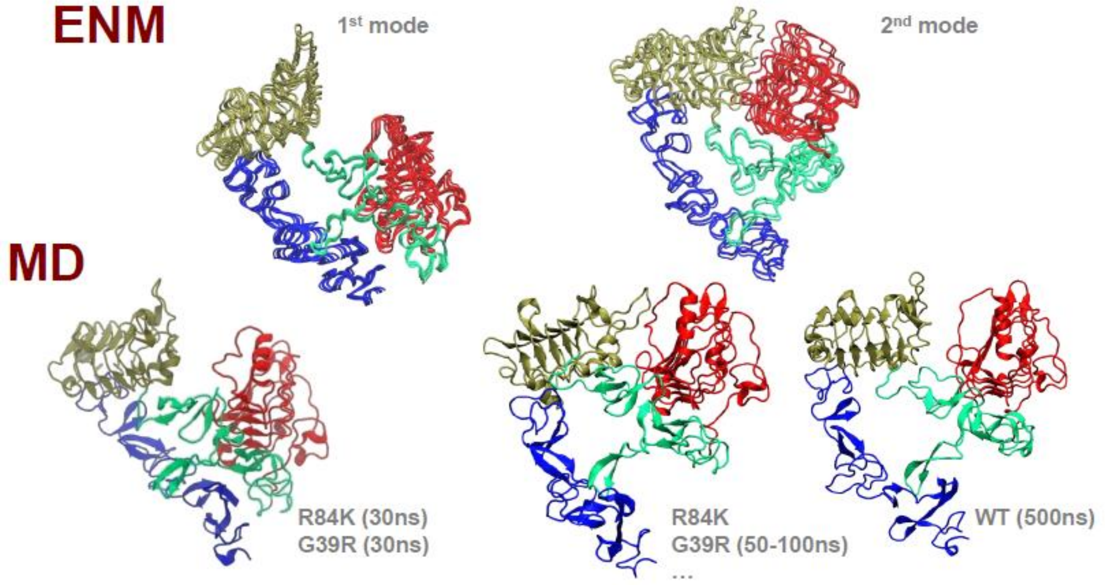


Figure 4. 14 **Spontaneous domain rearrangements in WT and R84K/G39R sEGFR follow the first normal modes.** Comparison between the two lowest-frequency modes of unbound closed sEGFR computed with ED-ENM (*above*) and unbiased MD samplings (*below*). Whereas WT-sEGFR can acquire a configuration with a closed binding cleft virtually identical to the 2nd normal mode after 500ns (*right*), the domain I-II R84K/G39R mutants sample this configuration in shorter timescales as well as the novel intermediate which clearly follows the 1st ED-ENM mode (*left*).

Our initial ENM analysis suggested that: i) the intrinsic motions of closed sEGFR are mainly relative rotations of the ligand binding domains, ii) this pattern of motions is altered by perturbations at the interdomain regions where mutations cluster, and iii) although the transition towards the open state is encoded in the intrinsic motions, it is not described by a single mode but requires a combination of several of them. Based on the ED-ENM screening results we focused on the R84K mutation at the I-II dynamical hot spot for MD simulations. The atomistic trajectories confirmed the ENM findings and demonstrated that, while the unbound ligand-binding domains spontaneously display rearrangements strikingly similar to the normal modes (*Figure 4. 14*) the transition towards an extended conformation is never observed in the absence of ligand. Apparently, the mutations at the I-II hinge interface act increasing the mobility of domain I, which enhances the sampling in the ns-timescale. Instead of the expected open untethered structure, R84K and nearby mutations at the I-II hinge interface can acquire a closed untethered configuration similar to the 1st mode. The biological significance of the observed intermediate, with an intriguing configuration different from known structures, is supported by a number of experimental evidences that do not fit with the open/closed conformers as we will discuss now.

First of all, the **tilting back of domain I exposes completely the Cys-bonded loop C287-C302**. As mentioned in the introduction, cancer-specific monoclonal antibodies mAb806 and mAb175 recognize an epitope in this Cys-linked module inaccessible in both close-tethered and open-untethered sEGFR (Gan et al., 2012; Johns et al., 2004; Jungbluth et al., 2003). Exposure of this epitope is the hallmark of EGFRvIII glioblastomas, where a deletion of the entire domain I leaves this region free. In fact, mAb806 and mAb175 were raised in mice against EGFRvIII and thus it came as a surprise that they can also bind WT EGFR. Further research demonstrated that mAb806 target what is called a "*cryptic epitope*", i.e. one that is only exposed in a transient conformation of the receptor. In the intermediate state found this region is fully accessible, making easy for docking and MD to obtain sEGFR-antibody complexes (see *Figure 4. 16B*) which agree with crystallographic and mutagenesis data (*Figure 4. 17*) (Chao, Cochran, & Wittrup, 2004; Garrett et al., 2009). Not surprising, the superposition of the R84K intermediate and EGFRvIII reveals that both configurations remove the sterical block by domain I (*Figure 4. 15)*. Recent work has shown that ectodomain mutations bind selectively the inhibitors targeting the so-called "inactive" configuration of the kinase (Vivanco et al., 2012). It is tempting to suggest that this removal of domain I steric restriction not only exposes the cryptic site but affects the configuration of the kinase in a particular way, triggering specific signalling pathways. This could explain why these positions are recurrently mutated in diverse tumors (CNS,

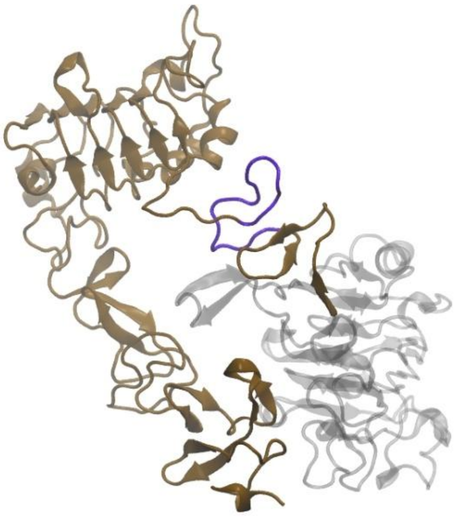


Figure 4. 15. **Alignment between the deletion mutant EGFRvIII and the R84K/G39R intermediate.**

EGFRvIII mutation (brown) is shown superimposed to R84K/G39R intermediate (light grey). The mAb806/175 epitope is shown in purple. Note how both states of the receptor remove the sterical blockage by domain I, exposing the cryptic epitope lung, colon or skin). Detailed sequencing of gliomas reveals that even different regions of the same tumor display multiple mutations belonging to this potential "functional" group - for example, EGFRvIII plus R84K (Joan Seoane, *personal communication*) - suggesting that tumor heterogeneity and mutation clustering could be the expression of a convergent evolution process.

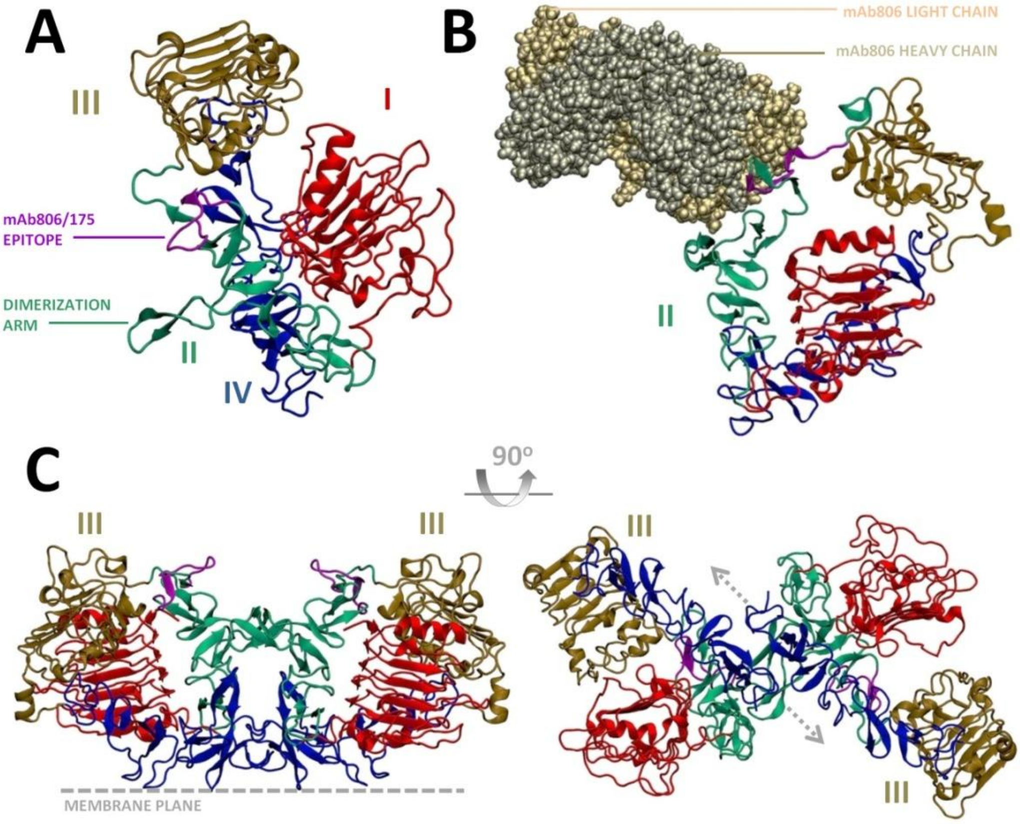


Figure 4. 16. **Ectodomain R84K untethered state.** A) **Domain disposition of sEGFR after the untethering transition.** Domains I-IV are colored red-green-brown-blue. Note how the extreme backward rotation of domain I, up to interact with domain IV, allows full access to the cryptic epitope (highlighted in purple). B) **Model of the complex of the antibody mAb806 bound to the R84K mutant intermediate state.** This figure shows one of the multiple possible models obtained after docking one of the intermediate snapshots to the crystal complex (PDB: 35GV), and performing a short MD simulation. During the simulation the epitope tends to extend making further contacts with the light chain. C) **Hypothetical "Crouched" unbound dimer of sEGFR**, resulting from direct docking through exposed dimerization arms from the same snapshot in A), front (*left*) and bottom view (*right*); note that the cryptic epitope (*purple*) remains exposed. The plane of the membrane passes under the structure in the front view (dashed line) and parallel to the paper in the bottom view *(not shown).* Note also the anti-parallel disposition of domain IV-IV, as well as the large domain III-III distance *(right)*.

Second, the **closed untethered intermediate seems perfectly designed to form a symmetric unbound dimer** through the interaction of the dimerization arms (see *Figure 4. 16C*, details in *Methods*). The resulting dimeric structure can be described as "*crouched*" or "*rod-shaped*", as opposed to the symmetric doubly-bound dimers formed by open monomers (compare with *Figure 4. 1B*), which have a *"proud"* or "*heart-shaped*" configuration. The existence of such alternative dimers is supported by a number of evidences which do not fit into the known crystal structures. The "*crouched*" dimer is still fully capable for the interaction with mAb806 *(Figure 4. 16C left*) (a feature of unbound dimers reported by (Gan et al., 2007)), and has an anti-parallel arrangement of IV domains (*Figure 4. 16C right*) in agreement with cross-linking and EM experiments (Mi et al., 2008). Similar domain IV arrangements have been observed in crystallographic closed dimers (Ferguson et al., 2003) and in long simulations of unbound and singly-ligated dimers (Arkhipov et al., 2013). A last piece of evidence supporting the existence of "*crouched"* dimeric structures comes from EM experiments (C. Lu, Mi, Walz, Springer, & Avenue, 2012) reporting unbound dimers with a distance between domains III-III of 118 ± 25 Å, close to the ones here found (115Å, see *Figure 4. 16C right, yellow stars*), but much larger than in X-Ray ligand-bound dimers (70-80Å). In summary, our simulations strongly suggest that WT sEGFR has a small but intrinsic tendency to spontaneously untether, exploring states that can form dimers different from the crystal ones but compatible with experimental data.

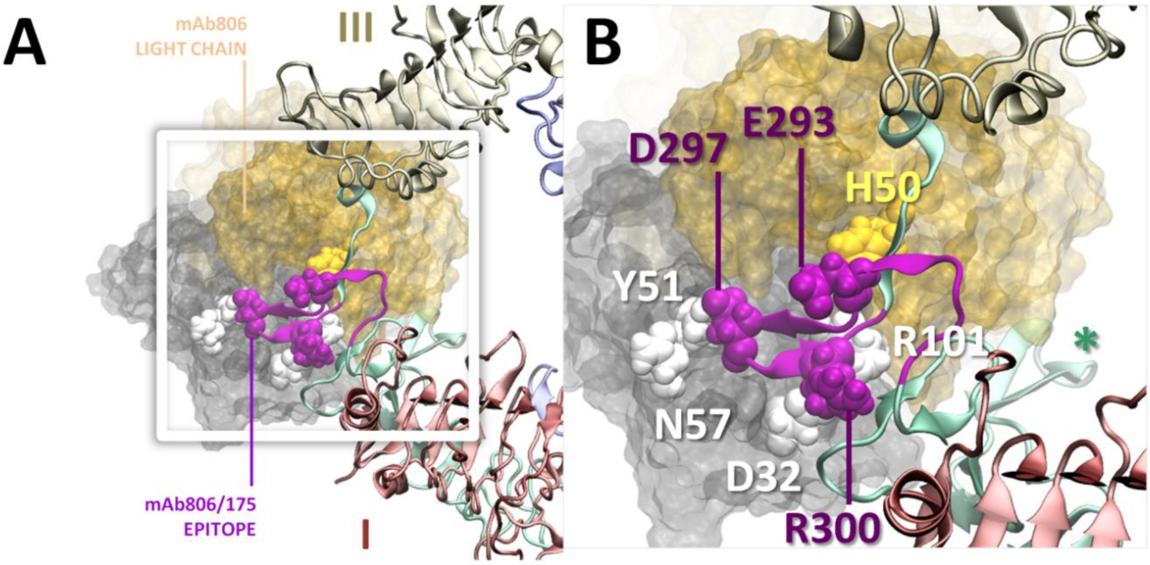


Figure 4. 17**. Detailed stereoview of the ma806 antigen-binding site and the R84K exposed epitope.**
Two levels of zooming showing the interaction between the mAb806 antibody and the R84K exposed epitope: (A) Overall view showing domain disposition and (B) Detail of the antigen-binding site. The ectodomain peptide backbone is shown as cartoon, with domains colored red (I), green (II) and tan (III), and the epitope loop (C287–C302) in magenta; the antibody is shown as surface with colored light (yellow) and heavy (gray) chains. Known interacting residues (E293-H50/R101, D297-Y51/N57, and R300-D32) are highlighted as white (heavy chain), yellow (light chain) or violet (EGFR epitope) spheres. Simple docking to the crystallographic complex (PDB: 35GV), followed by minimization to remove minor steric clashes, allows for close apposition of the main binding residues and the 806 binding site. This figure shows one of the multiple possible models in which the exposed C287–302 epitope can be inserted into the antigen-binding site; note the dimerization arm (asterisk), away from the binding region in (B).

Finally, **the intermediate found, characterized by a closed and compact but untethered configuration, is also compatible with SAXS experiment**s (Dawson et al., 2007), which unexpectedly revealed minor conformational changes upon tether weakening mutations, but not a transition towards the open state. Overall, our data suggest that ligand-independent untethering may not lead to extended, but to close-like configurations of sEGFR.

These results also confirm previous simulations of HER4 (Du et al., 2012) that indicate that the ligand is essential to drive the ectodomain to an extended, but still inactive configuration. Clearly, this suggests that an additional dimerization-mediated step is required to reach the fully active state, characterized by a “high-affinity” binding site. Our ENM and MD simulations also indicate that once reached, this active open structure is the most stable conformation of the receptor, and that not only the tethering motive, but also other interdomain contacts (such as the ones restricting domain I oscillations) contribute to keep unbound sEGFR in a tethered state – a self-inhibition mechanism bypassed by oncogenic mutations. Our simulations support then the hypothesis (Liu et al., 2012) that vertebrate sEGFR evolved from an open structure in invertebrates (Alvarado, Klein, & Lemmon, 2009, 2010), by creating molecular restraints to stabilize a tethered form and allow a more sophisticated regulation.

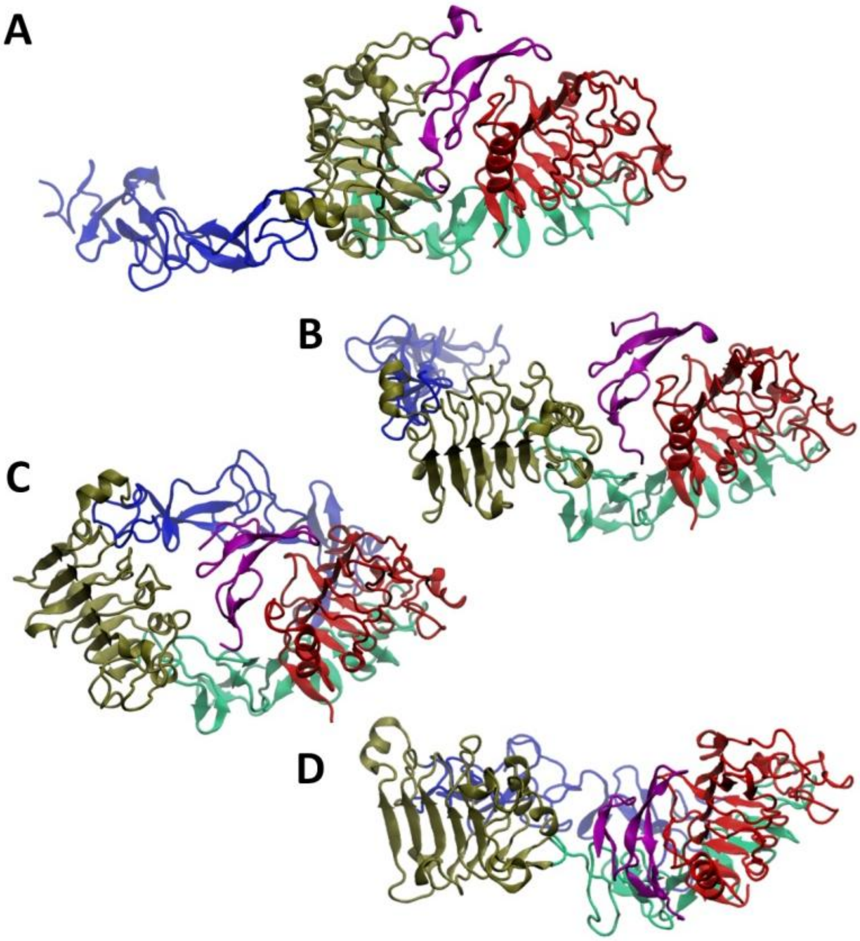


Figure 4. 18. **Domain I - III observed orientations for EGF binding (top view).**

Domains I-IV colored red-green-brown-blue, EGF in purple. Whereas EGF only binds domain I in the closed state (D), the smaller cleft in the intermediate state (C) and at extended state found at the end of WT + ligand simulations (B) allows for simultaneous binding to both surfaces at the same time. However, a large rotation of domain III is still needed to allow for high-affinity binding as observed in the active state (A).

Another point of interest highlighted by our simulations is the role of the ligand in the activation process. Ligand binding to EGFR displays a characteristic concave-up Scatchard-plot, related either to negative cooperativity or heterogeneity of binding sites, typically identified by two “high-affinity” and “low-affinity” classes corresponding to open and close states (Krall, Beyer, & MacBeath, 2011; Özcan, Klein, Lemmon, Lax, & Schlessinger, 2006). Although the role of negative cooperativity in the ectodomain is clear according to recent studies (Alvarado et al., 2010; Pike, 2012), our simulations also suggest a significant diversity of binding modes due to the changing arrangements of the ligand-binding domains. Whereas EGF binds always to the same site in domain I, domain III can present very different surfaces for binding in untethered states (see domain I-III orientations in the intermediate, and open inactive and active states, *Figure 4. 18*); in the tethered conformation, both domains can even collapse occluding the binding site, presumably leading to a low-affinity state. Such different configurations can clearly determine variable binding affinities and modulate the overall dynamics directing the motions of the flexible domain I. Our simulations show that first, low/intermediate affinity binding to domain I can drive receptor opening by an induced fit mechanism and suggest that after dimerization and activation, high affinity binding locks the domain I-II hinge and stabilizes the bound dimer.

To end this discussion we will address the potential significance of a ligand-independent untethering process. Recent data demonstrates that the bioactive species at the low EGF levels existing *in vivo* is an asymmetric singly-, not doubly-, ligated symmetric dimer (Alvarado et al., 2010; Liu et al., 2012) which would require higher ligand concentrations (such as that found in the crystal). Clearly, the formation of singly-ligated dimers requires the spontaneous untethering of one monomer. The similarity of WT and R84K spontaneous motions observed in our simulations strongly suggest that the WT ectodomain may also sample closed-untethered structures. It has been reported that the mAb806 epitope is exposed in a small population of WT untethered receptors, especially when glycosylation is altered – a condition which clearly can change domain I-II relative flexibility (Gan et al., 2012; Johns et al., 2005).

Such ligand-independent untethering mechanism (see *Figure 4. 19*) similar to the one described for the R84K mutation, running in parallel to ligand-driven untethering, would imply the existence of a small but basal subpopulation of untethered intermediates, as well as of pre-assembled unbound “flat” dimers (see *Figure 4. 19*). Whereas free untethered monomers could act as *ready-to-dimerize* partners for the ligand-untethered ones, the unbound pre-active dimers could also act as first target for ligand binding (as suggested by (Teramura et al., 2006)), forming in both cases asymmetric singly-ligated dimers *(Figure 4. 19)* at low EGF levels. The transition towards the fully active state would hide again the cryptic epitope, as reported by Walker et al. (Walker et al., 2004), who demonstrated that tether-weakening mutations increase mAb806-reactivity, but that this disappears as EGF binds. In the presence of oncogenic mutations, the populations of untethered monomers and unbound dimers would increase dramatically, enhancing basal kinase activity and EGF-response. Overall, the proposed model reveals the existence of multiple patterns for EGF-response, which could be sensitive to differential ligand concentrations or post-translational modifications.

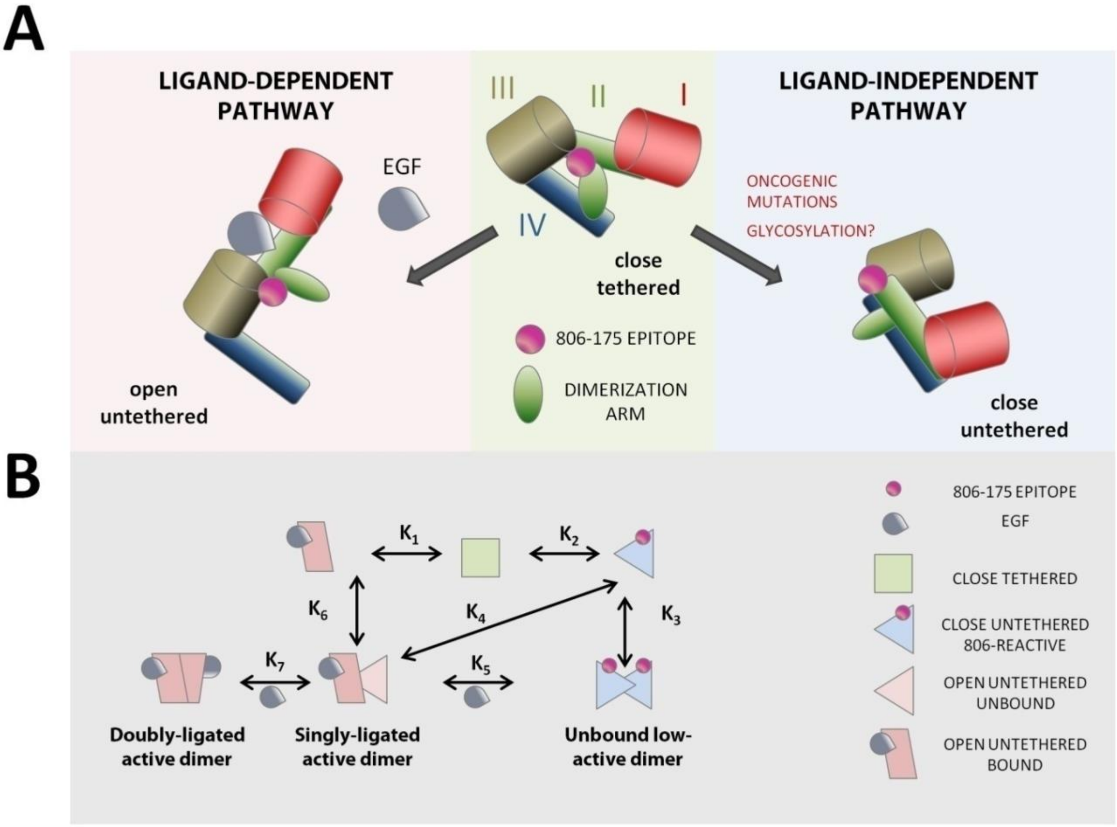


Figure 4. 19**. A novel model for EGFR activation.**

A) Proposed pathways for sEGFR untethering, showing domain I-III rearrangement and exposure of the dimerization arm (*green*) and the cryptic epitope 806/175 (*magenta*). **B) A model for sEGFR dimerization and its relationship with 806/175 epitope recognition.** Closed sEGFR is mostly closed (*green square*), but a small population untethers spontaneously into the 806/175 reactive-intermediate (*blue triangle*), specially under certain conditions changing local flexibility at the hot spots (mutations, glycosilation, etc.). At low, physiological EGF concentrations, this constitutive population of untethered intermediates can either form unbound dimers that react to EGF, or act as a ready-to-dimerize partner for EGF-untethered receptors (*pink rectangles*), forming the bioactive singly-ligated dimers, upon further conformational changes. At high EGF concentrations, the doubly-ligated dimers can be formed either from ligand binding to the singly-ligated ones or from two EGF-untethered monomers (*not shown*). All dimers shown represent an interaction between domain II-II dimerization arms; pre-dimerization through domain IV-IV interactions is excluded for the sake of clarity but could play an important role in an fast EGF response.

Table 4. 6 Summary of the MD simulations performed.

DESCRIPTION	DURATION (ns)	NUMBER OF REPLICAS	NUMBER OF ATOMS
SIMULATIONS CLOSED STATE			
Closed state (1NQL, chain A) WT, no ligand	500	1	180,834
	200	1	
	50	2	
Total simulated time closed WT (µs)		**0.8**	
Closed state (1NQL, chain A) R84K, no ligand	500	2	180,461
	50	2	
Total simulated time closed R84K (µs)		**1.1**	
Closed state (1NQL, chain A) WT + EGF (1NQL, chain B)	500	1	181,189
	200	1	
	50	2	
Total simulated time closed bound WT (µs)		**0.8**	
Total simulated time closed state (µs)		**2.7**	
SIMULATIONS OPEN STATE			
Open state (3NJP, chain A, residues 3-614) WT, no ligand	200	2	329,799
	50	2	
Total simulated time open WT (µs)		**0.3**	
Open state (3NJP, chain A, residues 3-614) R84K, no ligand	200	2	329,614
	50	2	
Total simulated time open R84K (µs)		**0.3**	
Open state (3NJP, chain A, residues 3-614) WT + EGF (3NJP, chain C)	200	2	330,379
	50	2	
Total simulated time open bound WT (µs)		**0.3**	
Total simulated time open state(µs)		**0.9**	
SIMULATIONS INTERMEDIATE STATE[&]			
Intermediate state, R84K, no ligand	50	5	136,374
Intermediate state, R84K, + EGF[$]	50	5	162,563
Total simulated time R84K intermediate (µs)		**0.5**	
Intermediate state, R84K, II-II docked dimer[#]	50	1	442,127
Intermediate state, R84K +mAb806 docked antibody[1]	50	1	225,483
Intermediate state, R84K +mAb806 docked antibody[2]	50	1	299,587
Intermediate state, R84K + mAb175 docked antibody[3]	50	1	300,146
Total simulated time R84Kintermediate complexes (µs)		**0.2**	
Intermediate state, WT, no ligand	50	5	136,208
Intermediate state, WT, + EGF[$]	50	5	162,670
Total simulated time intermediate WT (µs)		**0.5**	
Total simulated time intermediate state (µs)		**1.2**	
Total simulated time closed + open + intermediate states (µs)		**4.8**	

[&] ***Intermediate state shown in Fig.6A (representative frame from the R84K 500ns trajectory),*** [$] ***Intermediate state with EGF docked to domain I as in 1NQL,*** [#] ***Crouched dimer shown in Fig.6B-C, obtained by docking the intermediates through dimerization arms as in the 3NJP open dimer,*** [1] ***Intermediate state epitope docked to mAb806 as in model 2EXQ,*** [2] ***Intermediate state epitope docked to mAb806 as in crystal 3G5V,*** [3] ***Intermediate state epitope docked to mAb175 as in crystal 3G5Y.*** All computations done in this work were performed on a 64 Intel(R) Xeon(R) CPU E5-2670, 2.60GHz from a 768-machine (48x16cores) HP cluster (parallel computer), totaling ~ 60 years of cpu time.

4.6 Summary

Present simulations outline a complex mechanism for sEGFR activation that reconciles the physics of the system with functional evidences, still unexplained by current structural models. The starting NMA study suggests that the ectodomain has intrinsic large-scale motions of the different domains, which are controlled by strategic hinge sites and interdomain contacts. We develop a simple local perturbation algorithm, designed to detect critical regions for the conformational dynamics. Our study reveals how cancer-related mutations cluster in these interdomain hinge points and loops that control the biologically relevant motions, outlining a new dynamics-based mechanism of oncogenicity.

Molecular Dynamics confirms that wild-type HER possesses an extraordinary flexibility, exploring wide regions of the conformational landscape even in the ns timescale and spontaneously reorienting the main subdomains as predicted by NMA. Simulations of the known activator mutation R84K further confirm the dynamical impact predicted by the Normal Modes perturbation and prove that mutations can shift the large-scale motions and untether the receptor, supporting the coexistence of ligand dependent and independent untethering pathways. The close untethered state found, which exposes the mAb806 epitope, helps to rationalize a large amount of structural and functional data and may play a major role in EGFR activation. Overall, our findings unveil a previously unknown connection between ectodomain missense mutations, its immunogenic properties in cancer and its activation mechanism, which clearly can open novel avenues for basic and clinical research. On a purely theoretical level, the surprising correlations found between the MD and NMA and the spontaneous transitions observed, plus the coincidence between the predicted critical sites with oncogenic mutations, strongly suggest that some multidomain proteins explore a highly harmonical energy landscape, and that this intrinsic structure-encoded equilibrium dynamics can be regulated by specific interfacial contacts conserved during evolution.

4.6 Publications from this Chapter

Orellana L., Hospital A. and Orozco M. (2014) **Unraveling the dynamics of epidermal growth factor receptor through oncogenic mutations**. *JACS (submitted)*

“We all behave like Maxwell’s demon. Organisms organize.”

James Gleick

5- Further topics: from local to global conformational changes

This page intentionally left blank

Chapter 5 Further topics: from global to local conformational changes and network theory

In this chapter, we will present different applications of the ED-ENM method developed in **Chapter 3** to sample both local motions and large-scale conformational changes. As we have discussed throughout this thesis, proteins undergo large structural transitions to perform different functions, as demonstrated by open/close and bound/unbound pairs often observed by X-ray crystallography; an extreme example is the dramatic domain rearrangement of the EGF receptor seen in **Chapter 4**. Most of these conformational changes are encoded in the structure and intrinsic topology of each protein, and can be easily traced by one or two normal modes that provide precious information on the direction of functional transitions. However, if we want to analyze the detailed pathways for conformational changes, it is necessary to reconstruct some sort of ensemble or pseudo-trajectory to extract all the information contained in the normal modes and investigate the possible routes. In the next sections, we describe an ENM-derived algorithm to trace large transitions with minimal information from the target structure (Orellana, Carrillo, & Orozco, 2014), which can be applied to analyze large rearrangements. We also describe briefly an application of ED-ENM modes to guide the sampling of the conformational landscape by a coarse-grained model based on pseudo-physical potentials (Sfriso et al., 2012). A variation of the same method is used to generate ensembles from the ED or NMA principal motions, in order to analyze long-range concerted motions across beta-sheet motifs (Fenwick, Orellana, Esteban-Martín, Orozco, & Salvatella, 2013). Finally, the link between protein dynamics and intrinsic network properties is explored by a fast internal coordinates NMA (Orellana, Lopéz-Blanco, Chacón, & Orozco, 2014).

5.1 From the analytical to the numerical solution of the harmonic equation: Langevin-driven ensembles

An elegant alternative to unfold the information contained in the harmonic equation (*Eq.2.9*) is, instead of solving analytically by matrix diagonalization, solving numerically by performing a ***Brownian Dynamics (BD)*** simulation (J.A. McCammon & Harvey, 1987). Just three years after the publication of Einstein's description of Brownian motion, Paul Langevin modelled the movement of particles in a fluid with Newton's second law (Langevin, 1908). The protein is in a stochastic bath that keeps temperature constant, and the equation of motion for each residue *i*, represented by the coordinates of its C-alpha carbon (r_i), is given by the so-called Langevin equation:

$$m_i \ddot{r_i} = F_i - \gamma \dot{r_i} + \xi_i(t)$$ (Equation 5. 1 *Langevin Equation*)

The second term of the above equation is a dispersive force, accounting for the viscous resistance the particle feels on going through the fluid (depending on a friction coefficient γ), whereas $\xi_i(t)$ is a Gaussian noise term, due to the molecular-thermal agitation of the surrounding solvent, which leads to random collisions on the particle. The force acting on each residue *i*, F_i, is computed assuming harmonic potentials for its interactions with the rest of residues *j* as in *Eq. 2.16*, with the corresponding spring constants defined by the ED-ENM algorithm, which couples strongly the first chain neighbours, as described in **Chapter 3**. To solve the stochastic differential *Eq.5.1*, the Verlet Algorithm can be used to integrate numerically velocities from positions (as in (Carrillo et al., 2012)). Using this simple approach, different pseudotrajectories for the ENMs can be obtained changing the random seed of the algorithm each time. In this scheme, it is possible to further bias the stochastic dynamics to generate, for example, ensembles reproducing the normal modes. An additional force along the coordinates of the *m*-normal modes or principal components can be introduced, so that two forces act on each residue *i*:

$$F_i = \sum_k^m \frac{k_B T e_k}{\lambda_k} (r - r_0)\, e_i + F_i^*$$ (Equation 5. 2 *Modes Biasing Force*)

where e_k and λ_k are the eigenvectors and eigenvalues defining the external perturbation force due to the first *k*-principal components (or normal modes), and F_i^* is the force due to the harmonic interactions mimicking the internal covalent and non-covalent short-range forces computed from ED-ENM, acting here as a constraint *SHAKE*-like potential to keep the secondary structures intact. This approach is used in *section 5.3* to create a conformational ensemble that follows the essential motions and reproduces correlated motions in beta-sheets (Fenwick et al., 2013), under revision).

Alternatively, instead of an external force, a biasing ***Dynamic Importance Sampling*** **(DIMS)** algorithm based on an informational criterion (Beckstein, Denning, Perilla, & Woolf, 2009; Woolf, 1998; Zuckerman & Woolf, 1999, 2000), acting as a *Maxwell demon*, can be applied to generate transition pathways, for example, only accepting the moves that approach the initial structure towards the known target, as described in *section 5.2* ((Orellana, Carrillo, et al., 2014), in preparation)).

5.2 Large conformational changes: Exploring transition pathways with elastic networks

Proteins function as dynamic molecular machines that cycle between different states, changing in response to temperature, electrochemical gradients or presence of ligands. There are many examples of structures trapped in different conformations due to binding to diverse molecules, changes in pH or mutations, to name a few. However, since dynamic high-resolution techniques are still limited to small systems, the detailed mechanisms driving large-scale conformational changes are unattainable by experiments. Typically, very limited structural information is available for the intermediate conformations along a transition pathway. In conventional equilibrium MD, protein spends most of the time moving in a local minimum and only rare fluctuations allow overcoming free-energy barriers to access other states. The elucidation of the mechanisms behind conformational changes is difficult due to the multiple paths accessible and the transient nature of the intermediates involved. In systems with thousands of atoms, conformational changes are rarely observed – with exceptions such as the mutant EGFR simulations in **Chapter 4**. In spite of novel methods to force the sampling along the direction of the transition, MD simulations of large transitions are still very expensive computationally, as discussed in **Chapter 2**,. Therefore, CG approaches are a useful alternative and, as we also saw in **Chapter 4**, can help to define initial pathways for further exploration by atomistic MD. Among CG methods, ENMs are extremely powerful to predict with striking accuracy experimental conformational changes (F Tama & Sanejouand, 2001). More than 95% of the transitions in the MolMov database (Alexandrov et al., 2005) can be described by just a

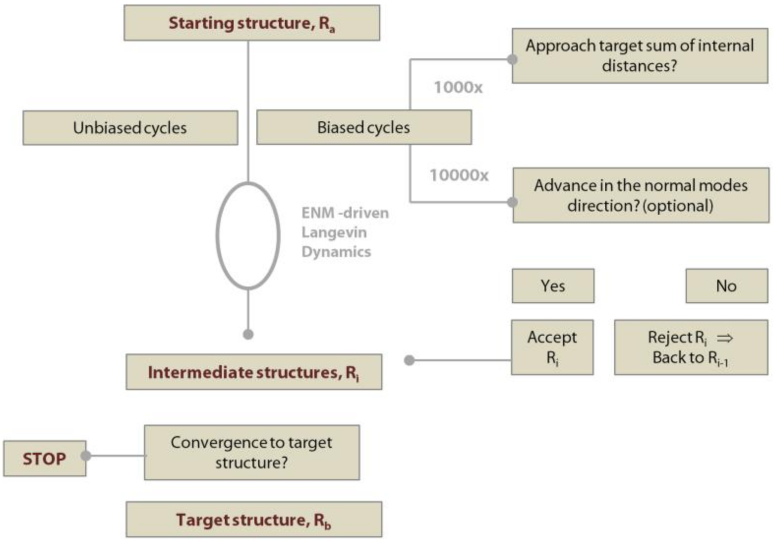

Starting structure, R_a
Approach target sum of internal distances?
1000x
Unbiased cycles
Biased cycles
10000x
Advance in the normal modes direction? (optional)
ENM -driven Langevin Dynamics
Yes
No
Accept R_i
Reject R_i ⇒ Back to R_{i-1}
Intermediate structures, R_i
STOP
Convergence to target structure?
Target structure, R_b

Figure 5. 1 Flowchart of the basic hybrid ENM-BD algorithm.

Biasing towards the target structure is introduced comparing with the sum of internal distances or following the normal modes which describe the transition.

couple of low-frequency modes from ENM, which therefore provide an excellent initial approach to sample large-scale conformational changes. We have seen an outstanding example for the large biomolecular machine EGFR in **Chapter 4**, where the MD samplings clearly follow the ED-ENM normal modes.

Usually, algorithms based on iterative NMA are used to trace transition pathways: the low-frequency modes are used to perturb the structure and generate new conformations from which to seed the next iteration step; this kind of approach is often used for building atomic models based on electron microscopy density maps (Kawabata, 2008; Suhre et al., 2006; Florence Tama et al., 2004). The computed paths are more realistic than those obtained by simple interpolation schemes, but also present major drawbacks: the eigenvalues usually render unrealistic amplitudes, and the lack of physical restraints produce stereochemically distorted models that require further energy minimization steps. These problems can partially be solved recomputing the lower modes for intermediate structures along the transition (Z. Yang, Májek, & Bahar, 2009), or performing movements in the EN-NMA internal coordinate space (Lopéz-Blanco et al., 2011). However, the trajectories obtained are still lineal ones and cannot account for random, short-frequency movements due to thermal noise.

Elastic Network-driven transition pathways

In order to overcome some of the above mentioned limitations of ENM sampling of transition pathways, we have developed a novel approach based on the implementation of the MD-derived ENM potentials (Orellana et al., 2010) in a Langevin Dynamics scheme (Carrillo et al., 2012) as presented in the former section. Instead of perturbing each structure along the lowest frequency modes, we use the EN potential to drive a Langevin simulation. In our approach, bond lengths are kept realistic not only by the harmonic potentials, but also by projection of the random velocities in internal coordinates at each step. Efficient biasing of the trajectory in the direction of the transition is achieved using minimal information from the target structure in a DIMS-based scheme (see *section 5.1*). Every certain number (*K*) of unbiased cycles, the sum of internal distances (the progress variable, Γ_i) for the instantaneous structure, R_i, is recomputed and compared with the target one (Γ_t) and according to a Metropolis Monte Carlo procedure, used as criteria to accept or reject the proposed random moves (see *Figure 5. 1*). An additional bias to avoid unphysical paths can be introduced by explicitly sampling along the ED-ENM encoded direction, selecting the random moves that overlap with different combinations of the softest modes. Note that in either case the biasing along the normal modes is not directly driven by a force as in

Equation 5.3. According with the MoDEL database (Meyer et al., 2010; Rueda, Ferrer-Costa, et al., 2007c), 2-4 Å (depending on system size) is a good estimate of the oscillation around equilibrium structures generated by thermal noise; therefore, we assume that convergence into the target basin occurs when the sampled structures reach an rMSD with the target within this range, stopping the iteration cycle. Contrary to current ENM-based approaches, our method allows obtaining different ensembles along multiple transition pathways. To generate slightly different trajectories the combination and number of biasing normal modes (that are also chosen in different ways) can be changed in each simulation run. The moves following a given combination of modes are selected every 10000 cycles computing a simple overlap between the instantaneous transition vector ($\Delta R=R_i-R_B$) and the normal modes n-subspace considered V_n (comprising the n-modes chosen) (see Hess' metrics, *Eq. 3.2*). Furthermore, this EN-based approach allows easily exploring the role of specific contacts or particular structural regions in the conformational changes, by introducing perturbations in the ENM topology matrix as seen in **Chapter 4**. Since the particular moves in the Langevin scheme depend on the starting random seed, further variability can be introduced by changing it at each run.

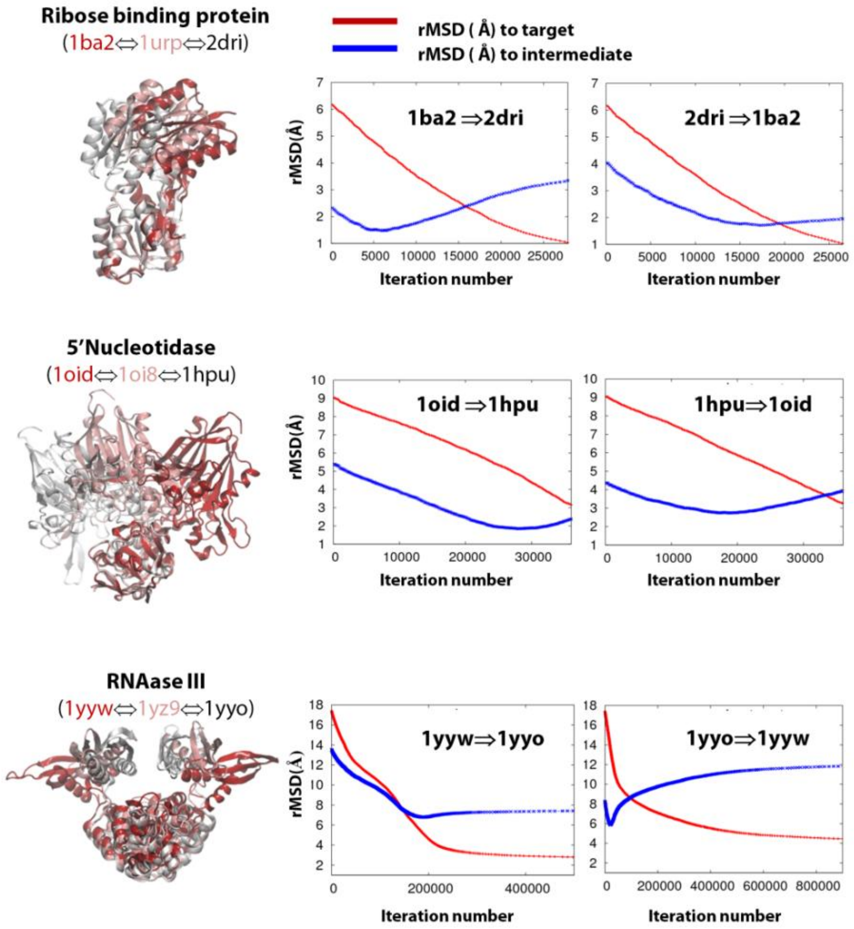


Figure 5. 2 **Evolution of the rMSD to the target and the experimental intermediate along the simulations in both directions.**

Three examples of proteins for which an experimental intermediate in a close to open conformational change is available are shown.

An excellent validation of any method to sample conformational transitions is the comparison of the trajectories obtained with the path followed by proteins undergoing large conformational changes for which there is a transition intermediate solved (Weiss & Levitt, 2009). Inspection of the rMSD profiles along our simulations (see *Figure 5. 2*) for some known cases shows that the algorithm can approach the natural pathway for the transition within an rMSD below 2Å in just minutes. In more challenging transitions, the method also shows an efficient performance, reaching the target structure in a few hours. The different trajectories obtained in the forward and reverse transitions demonstrate that in general, it is easiest to compute trajectories starting from the open state (that generally displays higher overlaps with the normal modes, see for example (F Tama & Sanejouand, 2001)). As mentioned before, stereochemistry is kept by the ENM potentials as well as an internal coordinates projection of the modes, so that further reconstruction of the backbone using the software PULCHRA (Rotkiewicz & Skolnick, 2008) renders intermediate structures with good geometries according to energy scores such as PROSA (Wiederstein & Sippl, 2007) (See *Figure 5. 3*).

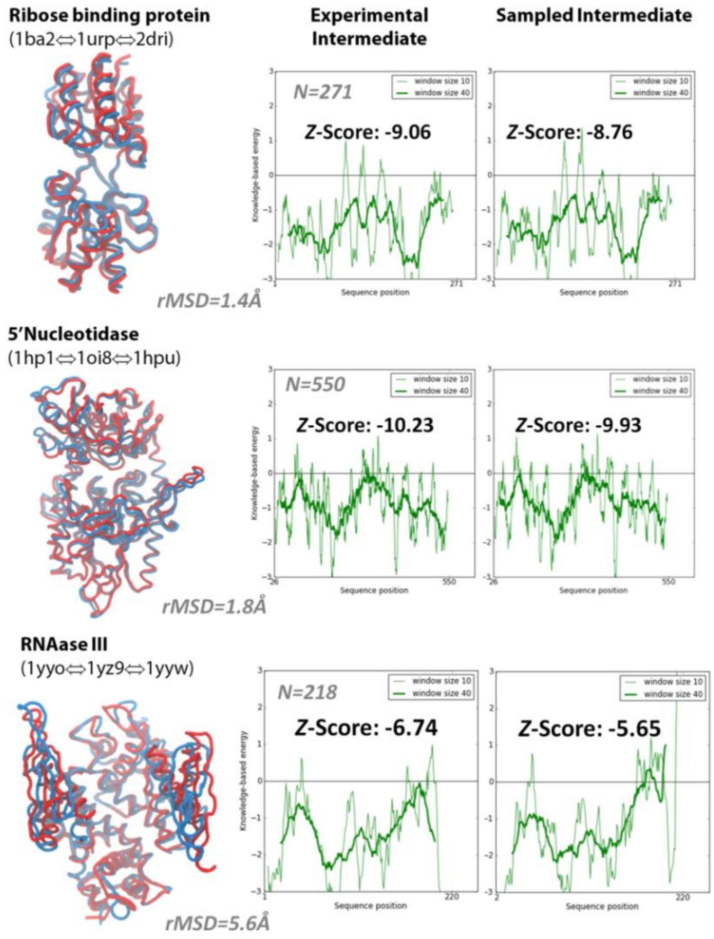


Figure 5. 3 Similarity between experimental and ENM-BD transition intermediates.
Left: C-alpha trace alignment between the experimental intermediate and the closest ENM-BD sampled structure. Right: PROSA profiles for the experimental and ENM-BD sampled intermediate.

Normal modes to sample the conformational space

An alternative but very powerful method to perform fast atomistic simulations is, instead of just coarse-graining the structure, to use pseudo-physical potentials, as discussed in **Chapter 2**. These techniques, in combination with normal modes biasing, can accelerate the sampling of conformational space extremely. In a recent work (Sfriso et al., 2012), the combination of ***discrete Molecular Dynamics*** (dMD, see **Chapter 2**) with biasing techniques based on normal modes and Maxwell–Demon sampling allowed a highly efficient exploration of transition pathways at the atomistic level. The core of this morphing procedure is the biasing algorithm, which enhances dMD sampling in the direction of the transition and also along the ED-ENM normal modes (*Figure 5. 4*). The good agreement between coarse-grained methods and MD (Emperador et al., 2008) guarantees the physical consistency of this hybrid ENM/dMD approach. The method follows a biasing Maxwell–Demon approach similar to that described above. Here, after a certain simulation step (*t*), a progress variable (Γ) is computed and compared with that at the previous accepted movement (*t* – Δ*t*), and accepted or not based on a probability p_t. To avoid that the purely informational criterion bias the transition to biologically unphysical paths, a second bias is introduced to guarantee that the transition follows the intrinsic flexibility of the protein. Whether a move follows the normal modes or not is again quantified by a simple overlap metric between the instantaneous transition vector and the subspace defined by the combination of ED-ENM eigenvectors that better reconstructs the structural change.

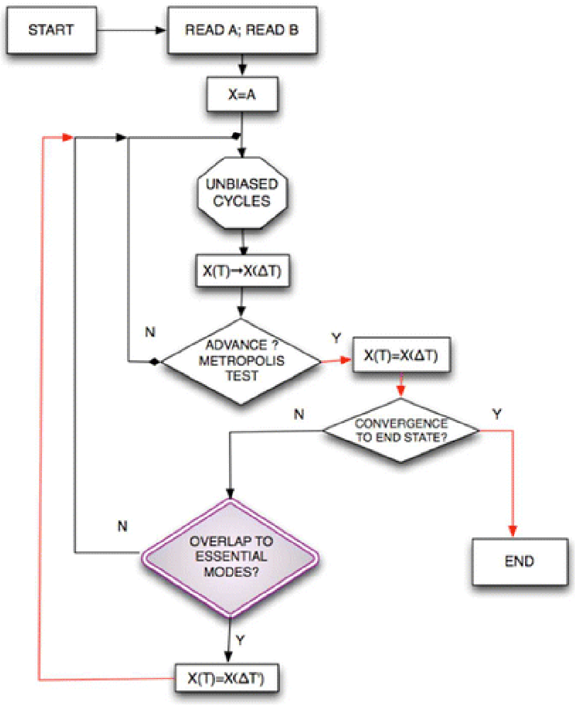

START
READ A; READ B
X=A
UNBIASED CYCLES
X(T)→X(ΔT)
N
ADVANCE ? METROPOLIS TEST
Y
X(T)=X(ΔT)
N
CONVERGENCE TO END STATE?
Y
N
OVERLAP TO ESSENTIAL MODES?
END
Y
X(T)=X(ΔT')

Figure 5. 4 **Sampling the conformational space with normal modes.**

Flowchart of the basic MDdMD method. Detail on the implementation of the NMA bias based on the initial and current overlap between transition and essential deformation space.

Published in: Pedro Sfriso; Agusti Emperador; Laura Orellana; Adam Hospital; Josep Lluis Gelpí; Modesto Orozco; *J. Chem. Theory Comput.* **2012,** 8, 4707-4718. DOI: 10.1021/ct300494q Copyright © 2012 American Chemical Society

5.3 Local conformational changes: correlated motions

Correlated motions in beta sheets are an inherent property of proteins

Current thinking on allostery predicts that correlated motions occur in all proteins (Gunasekaran, Ma, & Nussinov, 2004) and are the basis for information transfer and signal transduction. Although much work has therefore been directed to measuring correlations between distal sites in the conformational dynamics, their detection has remained elusive up to date. Most studies have focused on backbone correlations but side chains seem also to play a role in propagating conformational changes over long distances (Davis, Arendall, Richardson, & Richardson, 2006; DuBay, Geissler, & Bothma, 2011; Fraser et al., 2009). The existence of pathways of correlated motions via hydrogen bonds is strongly suggested from theoretical studies in allosteric enzymes that contain a central β-sheet, where weakly correlated motions link interaction sites (Tolonen et al., 2011) and store energy for binding and catalysis (Piazza & Sanejouand, 2008). Recent experimental progress comes mostly from NMR spectroscopy (Istomin, Gromiha, Vorov, Jacobs, & Livesay, 2008; Reif, Hennig, & Griesinger, 1997; Vögeli, Yao, & Bax, 2008), which has detected non-covalent correlations in the ubiquitin β- motif by high-resolution experiments (***ERNST* ensemble**) (Fenwick et al., 2011).

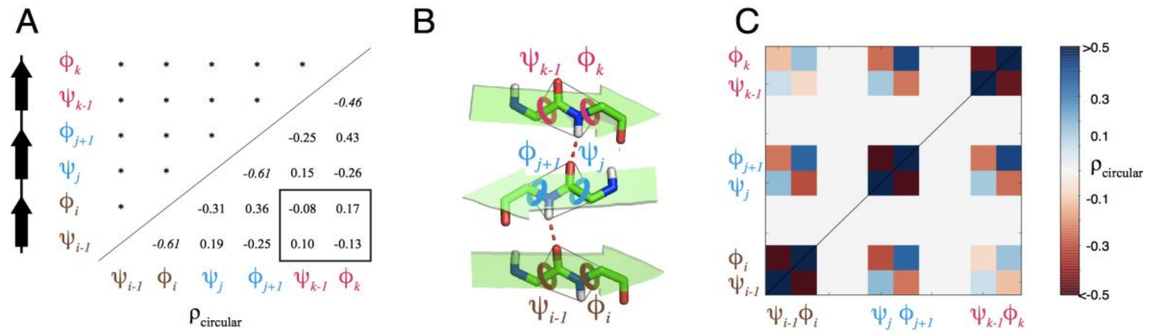


Figure 5. 5 **Circular correlation coefficients between dihedral angles show a checkerboard pattern.** β-sheets ensemble dihedral angle correlations coefficients (ρ, below the diagonal) observed within and between the strands (A) of the β-sheet motif (B). The dihedrals of the strands are indicated and labelled *i*, *j*, and *k*, corresponding to the three different strands as indicated in B. The long-distance correlations between strands *i* and *k* are highlighted with a box, while the crankshaft correlations are highlighted with italics. A graphical summary of the correlations is shown in C.

Here we describe an application of the BD algorithm presented in *section 5.1* to reveal how these important correlated movements are intrinsically encoded in the normal modes of the beta-sheet structures (Fenwick et al., 2013). We show that long-range concerted motions are a fundamental property of these structural motifs and that they are of functional relevance, first, by an analysis of the entire Protein Data Bank, and then, demonstrating how these correlations are intrinsic to the bending and twisting modes of beta-sheets and participate in signal transfer in allosteric enzymes.

An analysis of all the beta-sheets in the Protein Data Bank – considered as a collection of snapshots in different conformations - suggests the existence of these correlations as a general property in protein structures. To confirm this hypothesis and provide a rational explanation for this phenomenon, we analyzed BD/ENM-generated structural ensembles reproducing the normal modes and the principal components from MD. As noted in **Chapter 3**, both NMA and ED provide very similar descriptions of protein flexibility. The collective motions predicted from both methods show a remarkable agreement, and what is more significant, they contain the long-range correlations experimentally detected in the analysis of the Protein Data Bank.

The degree of correlation of the structural fluctuations of the different residues of the β-sheet motif can be evaluated by the ***circular correlation coefficient*** **(ρ)** for pairs of dihedral angles. These correlation coefficients are appropriate when both data come from circular or polar distributions (Jammalamadaka & Sengupta, 2001), i.e. of random variables whose values are angles in the range [0, 2π]:

$$\rho_{xy}^{circular} = \frac{\sum_{i=1}^{N} \sin(X_i - X).\sin(Y_i - Y)}{\sqrt{\sum_{i=1}^{N} \sin(X_i - X)^2 \sum_{i=1}^{N} \sin(Y_i - Y)^2}}$$

(Equation 5. 3 *Circular Correlation*)

the correlations were calculated for all the combinations of ϕ (phi) and ψ (psi) torsion angles in the β-sheet motif. The analysis of the entire protein data bank detected short-range correlations due to the crankshaft motion (*Figure 5. 5*, note values in the order of ρ ≈ -0.6) as well as strong short-distance non-sequential correlations between φ and ψ angles in neighboring β-strands caused by the β-lever. Most interesting, weak but significant long-range correlations between distal strands (*i=1, k=3*), similar to those observed in the ubiquitin ERNST ensemble, are also detected. To demonstrate that these correlations are shape-encoded in the topology of beta-sheet motifs, and therefore, an intrinsic property of these secondary structures, we studied the low-frequency motions in a benchmark of 24 β-sheet-rich proteins (see *Table 5. 1*).

Table 5. 1 Structure specific correlation analysis for the *minimal* beta motif.

PDB	Nres	Fold	Total Motifs	1PCA sig	1NMA Sig	5PCA sig	5NMA Sig
1a4h	214	2-layer sandwich	13	7	11	8	1
1cbs	137	beta barrel	19	2	1	15	8
1cpn	208	sandwich	17	5	4	7	7
1d2u	184	beta barrel	22	10	4	10	7
1g7n	131	beta barrel	18	7	6	6	8
1gbg	214	sandwich	21	7	4	11	8
1gnd	430	2/3-layer sandwiches	12	6	2	2	3
1gof	639	7-blade propeller	30	15	3	19	4
1icx	155	2-layer sandwich	17	8	3	9	3
1ifc	131	beta barrel	16	4	3	5	4
1ij9	196	sandwich	16	7	1	7	2
1msc	129	2-layer sandwich	9	5	5	6	2
1mvg	125	beta barrel	19	8	6	11	6
1ngl	179	beta barrel	12	1	1	5	2
1nkg	508	distorted sandwiches	31	12	3	18	6
1p6p	125	beta barrel	16	7	1	5	6
1plr	258	Box	25	9	7	12	2
1wp5	323	6-blade propeller	14	6	6	8	3
1yfq	342	7-blade propeller	21	8	5	8	6
2axf	385	1/2-layer sandwiches	18	7	5	9	8
2axg	385	1/2-layer sandwiches	18	10	7	9	11
2ayh	214	sandwich	23	10	1	16	6
2bvo	385	1/2-layer sandwiches	18	9	11	12	11
2cbr	136	beta barrel	18	5	4	10	4

Nres – the total number of residues in the structure.

Fold – the fold definition for the structure.

Total Motifs – the number of motifs that exist in the structure

1PCA sig. – the number of motifs that have scores greater than 10 for the ensemble of structures generated with Brownian dynamics using just the first PCA component.

1NMA sig. – the number of motifs that have scores greater than 10 for the ensemble of structures generated with Brownian dynamics using just the first NMA mode.

5PCA sig. – the number of motifs that have scores greater than 10 for the ensemble of structures generated with Brownian dynamics using just the first five PCA components.

5NMA sig. – the number of motifs that have scores greater than 10 for the ensemble of structures generated with Brownian dynamics using just the first five NMA modes.

These motions, encoded in the first modes, are known to correlate with large-scale, concerted motions, and are also in agreement with normal modes and experimentally described transitions. Using the mixed ENM-BD, we generated ensembles for the first MD-PCA components extracted from the MoDEL database or from the first ED-ENM Normal Modes, and later reconstructed with PULCHRA optimizing the hydrogen bonding geometry for the backbone atoms. The circular correlation coefficients are compared to those in the checkerboard pattern by a motif score which represents the number of correlations with the correct sign (with a maximum value of 15 and a minimum of 0); in *Figure 5. 6* the regions in which long-range correlations appear along the first mode are highlighted in red. As can be observed, motion along the first PC from MD/NMA does indeed give rise to the checkerboard pattern. This suggests that long-range correlations participate in collective motions associated to function. Interestingly, long-range correlations are not detected in μs simulations but become apparent after removal of the high-frequency motions by PCA, suggesting that current force-fields are not capturing the concerted motions in the simulated timescales.

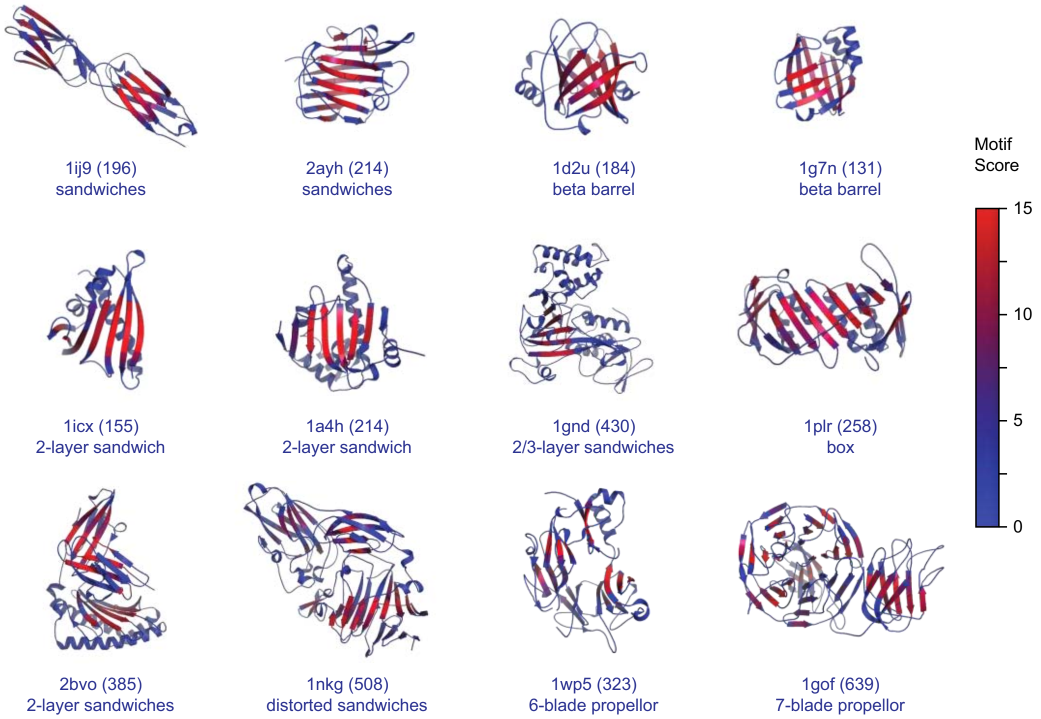


Figure 5. 6 **Beta-sheet correlated motions are intrinsically encoded in the principal components and normal modes.**

The β-sheets rich proteins with minimal motifs that behave as predicted from the model β-sheets. The structures are coloured according to the motif score (see text) and indicates that the motif moves in a correlated way as predicted from the PDB and the NMA analysis. The correlations were extracted from coarse-grained BD simulations that were generated using the 1st PCA mode from MoDEL MD simulations. We note that the results are invariant if more PCA modes are used in the Brownian dynamics (not shown), similar trends are observed if NMA modes are used instead.

Twisting and bending of beta-sheets drive allosteric transitions

To further determine if structural transitions from one state to another occur via functional changes that invoke correlated motions we studied a small subset of selected X-ray open/close conformers of beta-sheet proteins; the smallest ones (*1szv, 1s2h*) are single domains with a central β-motif and the rest are multi-domain structures (*1ram, 3dap, 1rkm*). In all cases we focused our analysis on the major beta-sheet motif present in each structure, defined as the largest displaying several high-scoring motifs. The global conformation changes are accompanied by significant local twisting or bending deformations of the major β-sheets between 0.5 - 2.5 Å associated with binding to other proteins, ligands or small molecules. We sampled the transitions from the open to the close conformer using Elastic Network-driven Brownian Dynamics as described in *section 5.1*. The β-sheet-rich regions of all five proteins produced the characteristic checkerboard pattern of the long-range correlations, and revealed how series of weak but long-range inter-strand correlations can create channels for the propagation of structural information from one extreme to the other in the primary β-sheets, spanning distances up to 20Å and thus connecting distant regions (see correlated residues highlighted as yellow balls in *Figure 5. 7*).

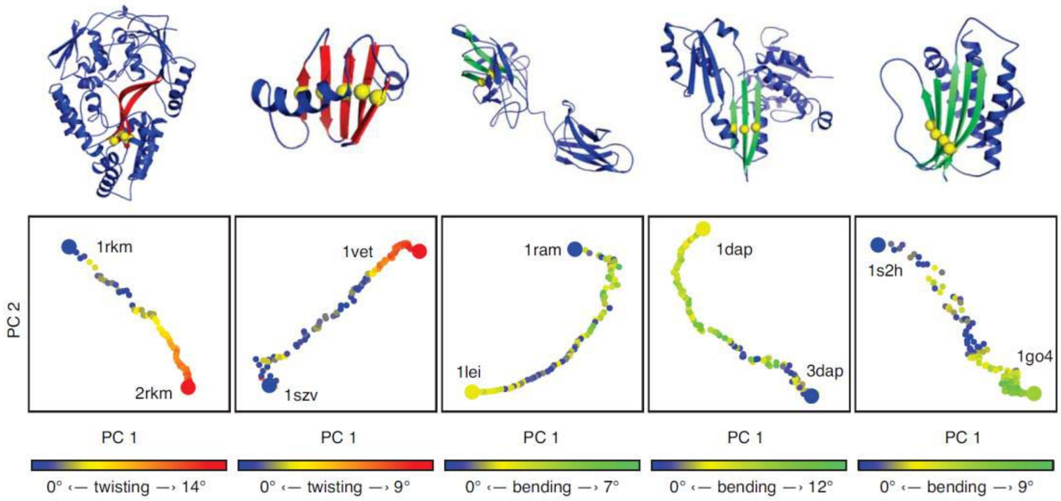


Figure 5. 7 **Correlated twisting and bending of beta-sheets drive large-scale conformational changes.** Brownian dynamics transition pathways projected onto the first two PCA components from unrestrained MD simulations. β-sheets rich regions are colored red or green to indicate either dominant twisting or dominant bending respectively while overlapping channels of motifs associated with the principal motif are shown in yellow. Transition pathways are shown and colored with red or green to indicate the dominant twisting or bending respectively during the structural transition. The color gradients indicate the change in degrees of the twist or bend from the start of the transition pathway. The two experimentally determined structures of the transition pathways are shown with larger point sizes. Pathways of signal transfer across the residues of the beta-sheet are highlighted as yellow balls.

To gain detailed insight of the role of the central β-sheets in these structural transitions, we used unrestrained MD simulations to further determine the stiffness of the bending and twisting motions and evaluate the deformation energies. For the set of five conformational transition pairs, we performed PCA on the MD simulations for both the full protein and the isolated primary β-motif present in each. The PCA of the entire structure yields a set of essential modes (Full-PCA) that describe the global conformational change; whereas the PCA of the reduced covariance matrix for the isolated β-motif (β-PCA) filters the pure bending and twisting modes which appear distributed in the first principal components of the entire protein. The structural transitions are related to significant bending (Θ_{bend}) or twisting (Θ_{twist}) of the β-sheet (2-10°), both in the crystal structures and the MD simulations. The eigenvalues of the first and second β-PCA modes scale with the number of strands as reported previously (Emberly, Mukhopadhyay, Tang, & Wingreen, 2004). These eigenvalues were used to obtain an estimate of the bending and twisting stiffness of the β-sheets via correlated torsions in the ns timescale, yielding values for beta-sheet bending on the order of 0.5-1 $k_BT/Å^2$ and as high as $3k_BT/Å^2$ for twisting (see *Table 5.2*).

In order to evaluate the potential energy stored in the correlated deformation of the β-sheets along complete functional transitions we used Mahalanobis distance for the first ten modes from the Full-PCA (Noy, Luque, & Orozco, 2008). This metric defines Euclidean distances weighted by the variance of every degree of freedom, which in the principal component orthogonal basis can be written as:

(Equation 5. 4 *Mahalanobis distance*)

$$d_M = \left[\sum_{i=1}^{n}\left(\frac{x_i}{\lambda_i^{1/2}}\right)^2\right]^{1/2}$$

Where x_i is the displacement along individual eigenvectors, λ_i stands for the corresponding eigenvalue (in units of distance^2), and the sum extends over the space of the first ten Full-PCA modes (*m*=10). The Mahalanobis distance represents the simplest deformation coordinate to drive a transition, assuming a harmonic relationship between displacements from the minimum and energy. Thus, in the harmonic limit the energy associated with displacements along principal components to reach the target structure can be easily determined from Mahalanobis distance as:

(Equation 5. 5 *Elastic Energy*)

$$E = \frac{k_b T}{2} d_M^2$$

As can be seen in *Table 5.2*, the first ten Full-PCA modes from the unrestrained MD simulations describe well the conformational change with overlaps of ~80%, being therefore possible to estimate the elastic energy of deformation along the conformational change from the Mahalanobis distance. For a conformational transition between two states, the minimal displacement along individual eigenvectors, x_i, is the projection of the transition vector, ΔR_{AB}, onto each PCA mode *i*:

$$x_i = \mathit{\Delta} R_{AB}.\cos\alpha = \frac{\mathit{\Delta} R_{AB}.v_i}{\|\mathit{\Delta} R_{AB}\|\|v_i\|}$$ (Equation 5. 6 *Langevin Equation*)

Where $\mathit{\Delta} R_{AB}$ is considered only for the subset of residues belonging to the primary β-sheet motif present in each structure. Projections for each one of the transition pathways onto the MD subspace defined by the first two principal components of the Full-PCA are shown in *Figure 5. 7* accompanied by the change in bending or twisting angle of the primary motif for each conformer along the transition pathway.

We found that the correlated bending or twisting of the beta-sheets can store energies in the range of 5-15 k_BT or ~0.02 kcal/mol/residue (see *Table 5.2*). These energy values are in agreement with previous calculations and suggest that twisting and bending may be carefully encoded in the structure of β-sheet rich proteins (Choe & Sun, 2007; Sun, Chandler, Dinner, & Oster, 2003). These additional examples show that the correlations that exist within sheet twisting and bending may define the routes for conformational transitions. One illustrative case is the large-scale conformational change found in the Periplasmic Binding Protein OppA, where the transition between the open and closed states involves the propagation of correlated bending and twisting motions of the β-sheet.

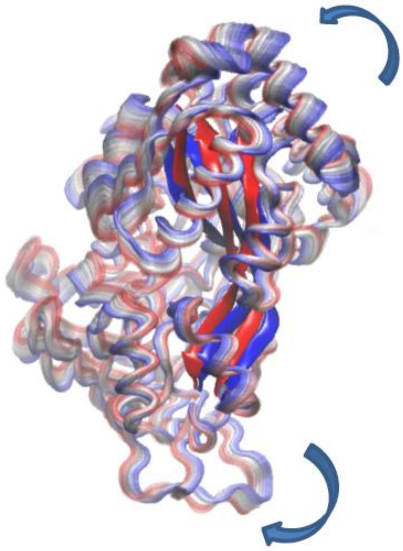


Figure 5. 8 **Correlated twisting of a beta-sheet in an interdomain hinge mediates a large open to close change in the protein OppA.**

The ligated (*1rkm*) and unligated (*2rkm*) forms of OppA are related by a rigid-body rotation of two structural domains. The hinge region that mediates the rotation is composed of two β-sheet segments at the interdomain interface; correlated changes in the backbone dihedral angles cause a twisting motion of the β-sheet that triggers the observed large-scale domain rearrangement (see *Figure 5.8*).

Table 5. 2 Twisting and bending of the major beta-sheet motif in large-scale conformational changes.

PDB	Nres	rMSD	Total Motifs	Sig	Θ_{bend} Θ_{twist}	Λ_{bend} Λ_{twist}	K_{bend} K_{twist}	Δ_{bend} Δ_{twist}	O_{10}	E_{10}
1szv/1vet	91	5.2	9	3	163±8 º 155±10 º	23.7 4.8	0.02 0.12	5º 10º	0.60	**4.1** **0.04**
1s2h/1go4	190	4.9	13	3	170±7 º 130±6 º	18.4 4.3	0.03 0.14	12º 2º	0.80	**6.7** **0.03**
1ram/1lei	273	3.0	5	2	170±5 º 146±8 º	0.8 0.6	0.74 0.98	3º 1º	0.94	**6.2** **0.02**
3dap/1dap	320	4.2	4	3	162±4 º 155±5 º	4.2 1.6	0.14 0.37	4º 2º	0.94	**1.1** **0.003**
1rkm/2rkm	517	3.3	7	4	160±5 º 150±7 º	0.9 0.4	0.66 1.48	3º 10º	0.96	**8.5** **0.02**

Nres – the total number of residues in the structure.

rMSD – root mean square deviation (Å) between open and closed conformations.

Fold – the fold definition for the structure.

Total Motifs – the number of motifs that exist in the structure.

Sig – the number of β-motifs that have motif scores > 10 for the ensemble of structures.

Strands/Nres - Number of strands and residues in the major β-sheet motif displaying correlated motions

ϑ_{twist} / ϑ_{bend} *- Bending and twisting angles in MD simulations for the major β-motif present in the structure (see definition above)*

λ_{twist} / λ_{bend} *- Bending and twisting eigenvalues ($Å^2$)in MD simulations for the major β-motif present in the structure*

K_{twist}/ K_{bend}*- Bending and twisting stiffness constants (kcal/mol. $Å^2$) in MD simulations for the major β-motif present in the structure*

Δ_{twist} / Δ_{bend} *- Change in bending and twisting angles associated to the conformational change for the major β-motif present in the structure between open and closed conformations*

O_{10}*– Overlap of the X-ray transition with MD first 10 PCA modes (Eq.10) for the deformation of the major β-motif along the transition coordinate.*

E_{10}*– Elastic energy (kcal/mol, Eq.13) for the deformation of the primary-motif along the first 10 PCA modes to approach the target structure.*

Overall, this suggests that the correlations accompanying sheet twisting and bending in the lowest-frequency modes may have a functional role in conformational transitions, acting as a mechanism to propagate information and store energy across protein backbones. We predict that the correlated motions described here are fundamental to the geometry of β-sheets but will also be present in other secondary structure elements where weak interactions are important. Our results have established that these β-sheet correlated motions provide mechanistic pathways that allow functional transitions and transfer of energy between distal sites and indicate that the local and long-range correlated motions described here, which are derived from the geometry of the β-sheets and of the requirement for maintaining the hydrogen bond patterns, can be crucial for protein function.

5.4 Complex systems theory and protein network description

In this section of the thesis, we explore the relation between the network properties of proteins and its dynamics, using the Elastic Network Model (ENM) approach. Being residue network topology the key determinant of the normal modes, it provides a quick view of the dynamical impact of any change in the residue interaction patterns. A highly efficient elastic network model in internal coordinates, *ic-ENM* (Lopéz-Blanco et al., 2011), has been used to explore the relationship between network topology and near-equilibrium motions of proteins. The extreme computational efficiency of ic-ENM allows us to explore the entire parameter space for different distance-dependent functions used to define the network of residue-residue interactions. Comparing the network-dependent ic-ENM deformation modes with those obtained from atomistic MD simulations, from the analysis of NMR ensembles, and from the analysis of biologically relevant conformational transitions, we found robust connections between network topology and protein flexibility, which guide us in the refinement of the ic-ENM model. The comparison among the near-equilibrium motions predicted by ENMs and the flexibility from MD simulations, X-ray and NMR conformers, shows a striking regular pattern for all examined proteins, which can be correlated to intrinsic properties of the network. These findings point out certain topological properties are a universal feature driving protein dynamics, and can provide a rational basis for ENM parameterization.

Protein motions are encoded by network topology

During the last years, the **theory of complex networks** has been applied to phenomena ranging from statistical mechanics or social sciences to systems biology (M. E. J. Newman, 2003; Strogatz, 2001). Eventually, almost any many-bodies physical system can be modeled as a network: the elements of the system are represented by nodes and the interactions between them by links; within that framework, network theory can explain global properties emerging from the local interactions – such as information transfer, communication processes or robustness to perturbations. Graph and network theories have been used to describe several properties of proteins, for example residue fluctuations from packing densities (Halle, 2002), or to analyze protein rigidity (Jacobs, Rader, Kuhn, & Thorpe, 2001) or allosteric communication processes (del Sol, Fujihashi, Amoros, & Nussinov, 2006; B. Ma, Tsai, Haliloğlu, & Nussinov, 2011). It is also well described the correlation between protein topology properties and the folding behavior of proteins (Ganesh Bagler & Sinha, 2007; Dokholyan, Li, Ding, & Shakhnovich, 2002; Vendruscolo, Dokholyan, Paci, & Karplus, 2002). However, the possible link between network properties and the *dynamical* behavior of proteins has not been investigated. As we have seen in **Chapters 3-4**, the motions of a protein derived from ENMs are mainly dependent on the protein shape, as defined by the topology matrix of residue-residue contacts. We found that normal modes depend strongly on the local backbone connectivity but at the same time, are modulated by long-range contacts (Orellana et al., 2010). This raises the following questions: *Why is there such dependence between the properties of the network and the dynamics? Are there some intrinsic topological features that can explain these observations?* To answer this, we have exhaustively explored the impact of network properties on protein collective motions, thanks to the use of an ultra-fast **ENM-NMA implementation in internal coordinates (ICs), iMod**, developed by Chacón and coworkers (Lopéz-Blanco et al., 2011). The ICs are defined by the canonical backbone dihedral angles, ϕ and ψ – with the exception of the first ϕ and the last ψ angles of the chains – whereas the remaining angles and all covalent bond lengths are fixed to keep the backbone structure. In this scheme, the Hamiltonian describing the potential energy to displace the protein network from its equilibrium conformation is as follows:

$$\boldsymbol{E = \sum_{i<j} K_{ij}\left(d_{ij} - d_{ij}^{0}\right)^{2} + s\sum_{\alpha}(\theta_{\alpha} - \theta_{\alpha}^{0})^{2}}$$

(Equation 5. 7 Hamiltonian in dihedral space)

The first term is the elastic energy due to the spring interactions between all residue pairs *i* and *j*, where K_{ij} is the spring or force constant connecting them, d_{ij} is the distance and the superindex *0* denotes the equilibrium conformation. The second term represents an extra-torsional stiffness, *s*, related to each dihedral angle, ϑ_{α}, in order to avoid the tip effect, i.e. irrational low-frequency modes caused by floppy regions (M. Lu, Poon, & Ma, 2006). Within the NMA approach (Case, 1994; Nobuhiro Go et al., 1983), the energy *E* is used to build the hessian matrix of second derivatives, *H*. Then, the protein motions are decomposed by solving the generalized eigenvalue problem:

$$\boldsymbol{He_{k} = \lambda_{K} T e_{k} \quad k=1 \ldots N}$$

(Equation 5. 8 Generalized Eigenvalue Problem)

Where λ_k is the eigenvalue associated with the *k*-th normal mode e_k, and T is the kinetic energy matrix of the system; the eigenvalues are related to the frequencies, ω_k, as $\lambda_k = (2\pi\omega_k)^2$. As in Cartesian NMA, the lowest frequency modes represent the directions of motion in the IC space, whereas the eigenvalues describe their amplitude of vibration after proper conversion. Note that the low frequency modes computed by NMA in internal or Cartesian coordinates (CCs) are almost identical, with deformation spaces displaying near perfect overlaps which demonstrate that CC modes correspond to dihedral angle motions (Kitao, Hayward, & Go, 1994). We use again a coarse-grained model where each residue is represented by the position of the C-alpha carbon (see details in (Lopéz-Blanco et al., 2011)).

Exploring the parameter space for distance-dependent functions

In order to perform a systematic exploration of the impact of the topological properties on the dynamics, we build a wide range of networks using a smooth sigmoid function of the distance, d_{ij}, between each residue (node) pair i and j to set the corresponding force constant, K_{ij} (the weighted link):

- if $d_{ij}<R_c$, then $$K_{ij} = \frac{C}{\left(1+\left(\frac{d_{ij}}{d_0}\right)^p\right)}$$

- if $d_{ij}\geq R_c$, then $K_{ij}= 0$ (Equation 5. 9 Distance-dependent test function)

This function is governed by four parameters:

I) **An equilibrium distance, d_0,** which sets the inflexion point of the function, and thus decides the extension of local versus long-range contacts; we considered values ranging from 1 to 9Å, about half to twice the physical value of the average shortest Cα-Cα distance linking two neighbor residues (3.8Å).

II) **A power term, p**, that determines the slope of the function around the inflexion point, ranging from 1 to 20: the increasing power term sets the rate at which long-range and short-range interactions are distinguished around the equilibrium point; at sufficiently higher *p* the exponential tends towards a discrete cutoff function.

III) **The constant C,** that determines the maximum stiffness of the links, and thus the magnitude of the eigenvalues, but does not change the variance profiles – i.e. the way in which variance is distributed across the different modes. For the sake of simplicity, we set a test force constant of 1 kcal/mol.Å^2.

IV) **The cutoff value (R_c)**, which defines the point in which the tails of the curves are set to zero. We explored values from 4 to 16 Å but will focus the discussion on the results obtained with an average standard value of 10 Å.

In *Figure 5. 9* the different families of curves, covering almost the entire distance-dependent functions space, are displayed together with some networks they give rise to. Note that each curve is related to a particular network structure, but two main variables decide most of the topology of the interactions: the distance at which neighbors and non-neighbors are distinguished, which is determined by the inflexion point d_0 (*Figure 5. 9, top*) (the greater it is, the larger is the number of long range contacts included), and the sharpness of this transition, which depends on the power term, *p*. At low *p* values, the transition from local to nonlocal is smooth, whereas higher *p* values yield cutoff-like functions (*Figure 5. 9, bottom*). At middle *p* values (5 to 8), more long-range contacts are included as d_0 increases, giving rise from close-neighbors-only to fully connected networks (*Figure 5. 9, top*). On the other side, when both *p* and d_0 are low, networks tend to be fully connected too. The increase in *p* value restricts the number of long-range contacts included, leading again to nearest-

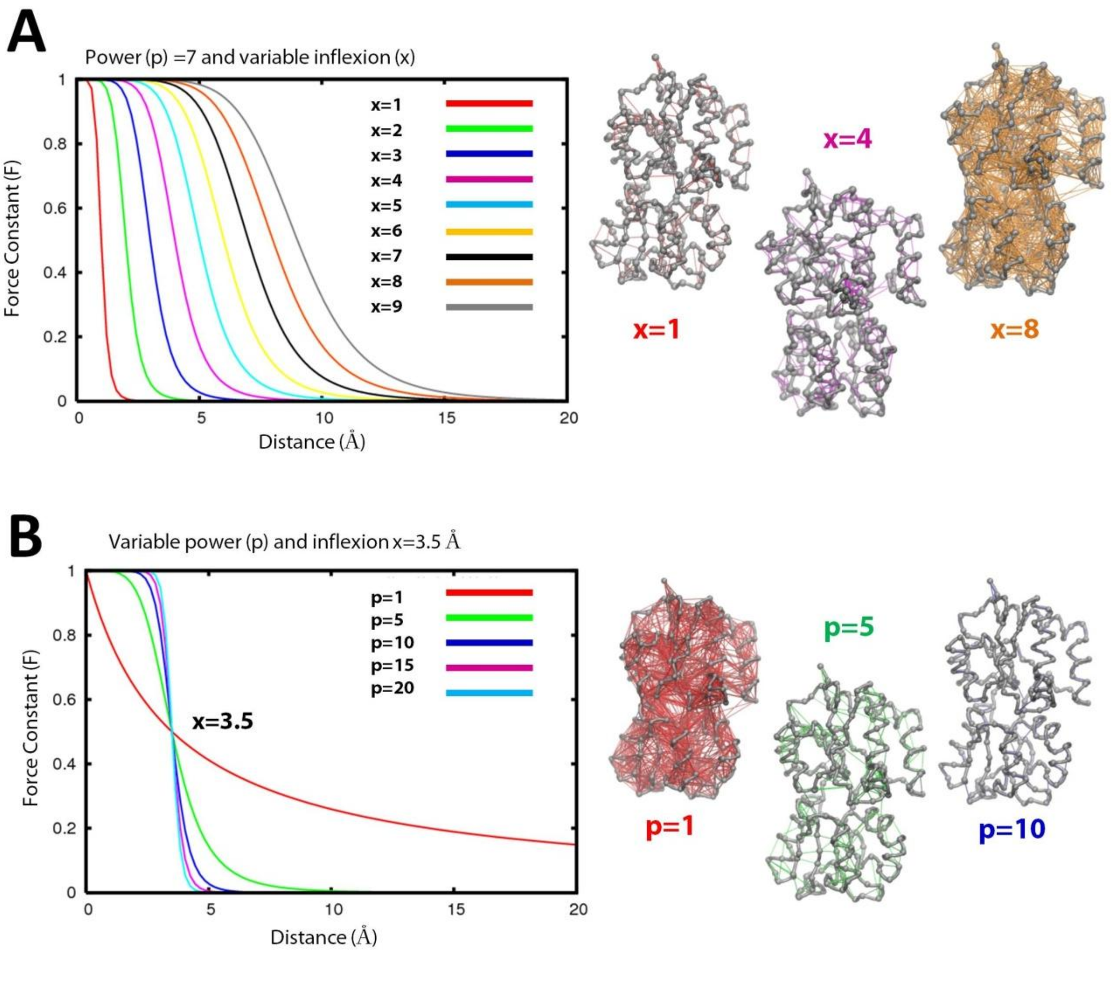


Figure 5. 9 Residue networks described by different force-constant function families
Different families of force constant functions (*left*), covering almost the entire distance-dependent space, are displayed together with some networks they give rise to (*right*) for a representative protein (*1lst*) at cutoff=10.

neighbors networks at low d_0 and high p networks (*Figure 5. 9, bottom*). Finally, the cutoff value R_c controls the length of the curve tail, i.e. the distance at which contacts are set to zero. At the wide cutoff range explored, the cutoff annihilates from close neighbors (4Å) to long-range contacts (16Å), with different impact on the network structure as determined by d_0 and p, as well as the network size set by N.

Network connectivity description

Here we consider a protein as a network composed of nodes (the protein residues) connected by edges (the spring constants) of strength dependent on the physical distance between them. There are a number of intrinsic properties that can be computed to analyze the structure of a network in terms of topology, i.e. the way nodes are connected to each other. In graph theory, the connectivity of a network is usually defined by the so-called *adjacency matrix*, A_{ij}, an NxN matrix composed by elements a_{ij}=1 if the nodes i and j are connected, and a_{ij}=0 otherwise; in the case of simple and undirected graphs, all diagonal elements are zero (a_{ii}=0) and this matrix is symmetric (a_{ij}=a_{ji}). Thus, by setting a distance cutoff, the topological structures of residue networks can be easily represented by the topology or Kirchhoff matrix, Γ_{ij}, which is equivalent to an adjacency matrix and can be analyzed using general network properties (Antal, Bode, & Csermely, 2009; G Bagler & Sinha, 2005; Di Paola, De Ruvo, Paci, Santoni, & Giuliani, 2012; Vishveshwara, Brinda, & Kannan, 2002).

However, a more realistic approach consists in substituting the binary topology matrix by a weighted graph (Barrat, Barthélemy, Pastor-Satorras, & Vespignani, 2004; M. E. J. Newman, 2004), assigning a value to each link. In a weighted network, a strongest connection between a pair of vertex/nodes/residues implies a greatest probability of contact or information transmission between them, and regarding flexibility, more correlated motions. Accordingly, we can naturally **consider the stiffness matrix, *K_{ij}*, defined in each point of the *(p, d)* space, as a weighted network graph**, instead of the typical 1/0 contact matrix. Here, the weight w_{ij} of an edge linking residues i and j represents the strength of the pseudo-bond coupling their motions, k_{ij}. As a consequence, all the metrics to analyze the network structure must be redefined in terms of a weighted distribution. We will consider the following network properties:

1) ***Average shortest path (characteristic) length, <l>, and average diameter <d>:*** in general, the distance between two nodes of a network is given by the length of the shortest path between them, l_{ij}, that is, the minimal number of edges (bonds) that need to be crossed to go from node i to node j (Dijkstra, 1959; M. E. Newman, 2001)**.** In an elastic network, the shortest/strongest the path the fastest the information transfer (correlation of motions) between them. Taking the average of

all the shortest paths , *l_{ij}*, which connect all residue pairs, we can compute an *average or characteristic length, <l>,* of the entire protein network:

$$L = \langle l \rangle = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} l_{ij}$$ (Equation 5. 10 Average Length)

which roughly measures the speed or efficiency of information diffusion between residues. From the shortest paths it is possible also to calculate the *average diameter, <d>,* of the network, i.e. the largest shortest path length between a pair of residues in the network.

2) ***Average clustering coefficient, <c>:*** The clustering coefficient measures the local group cohesiveness. In real networks, the existence of a link between nodes *i* and *j* and between nodes *i* and *k* often implies a link between nodes *j* and *k*. In other words, the clustering coefficient of a node *i, c_i*, measures the probability that its neighbors are also neighbors of each other; it equals 1 if a node is the center of a fully interconnected cluster, and equals 0 if the neighbors of a node are not connected to each other. In order to characterize the network as a whole, we consider the average clustering coefficient over all the nodes:

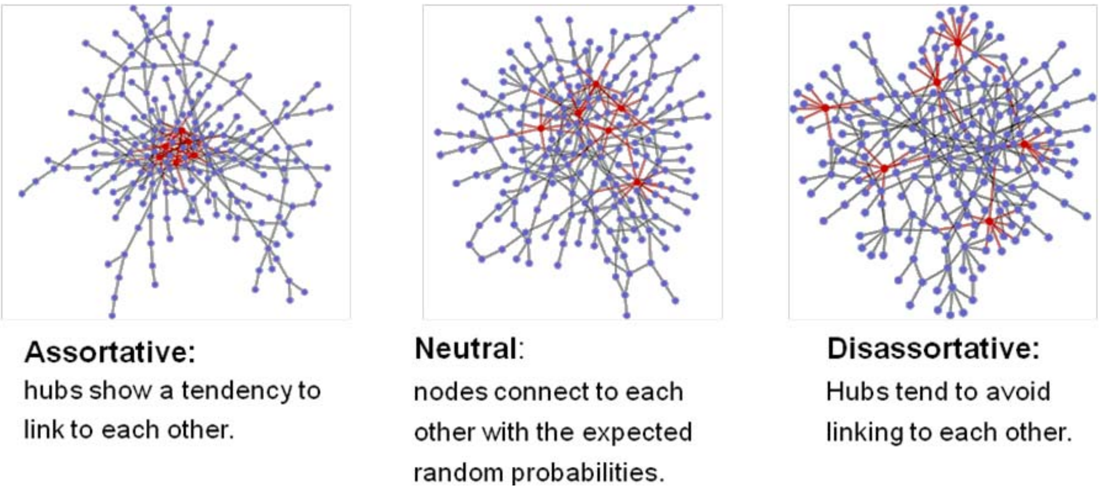
$$C = \langle c \rangle = \frac{1}{N} \sum_{i=1}^{N} \mathrm{c}_i$$ (Equation 5. 11 Average Length)

That expresses the cohesiveness, i.e. the global density of interconnected vertex triplets in the network. As before, we take into account a weighted definition of *c_i*, which considers the strength of the pseudobonds.

3) ***Average assortativity or assortative mixing.*** The coefficient of assortativity (M. E. J. Newman, 2002) represents the tendency of nodes to connect to others with a similar degree or number of connections. It is computed as the Pearson correlation coefficient (*r*) between the connectivity degrees of each pair of linked nodes, which takes values between −1 and 1. Positive values of *r* indicate a correlation between nodes of similar degree, while negative values indicate correlations between nodes of different degree. When *r* = 1, the network is said to have *perfect assortative mixing* patterns, while at *r* = −1 the network is completely *disassortative*. In terms

of protein structure, an assortative structure indicates a trend for hubs (highly connected residues, such as the ones in the protein core – inside globular domains, or structural elements) to connect preferentially to other hubs.

The above global connectivity properties of the residue weighted networks were computed using in-house developed code and the software *radatools:* http://deim.urv.cat/~sgomez/radatools.php developed by Arenas, Gomez *et al.* (Arenas, Fernandez, & Gomez, 2007; Gomez, Jensen, & Arenas, 2008). The correlation between the computed flexibility overlaps *(p, d)* maps and network properties *(p, d)* maps was measured by *Pearson Correlation coefficients*, which range from -1 (perfect negative correlation) to +1 (perfect positive correlation).



Assortative: hubs show a tendency to link to each other.

Neutral: nodes connect to each other with the expected random probabilities.

Disassortative: Hubs tend to avoid linking to each other.

Figure 5. 10 **Network assortativity measures the trend of a node to connect to others similar.**

In a assortative network (*left*) the highly connected nodes or hubs connect to other hubs, whereas in a disassortative network they connect preferentially connect to low-degree nodes (*right*)

The similarity between ENMs and protein flexibility depends on network properties

We performed ENM NMA calculations in internal coordinates for the same MD, X-ray and NMR protein benchmarks described in **Chapter 3,** but extended including longer MD simulations, as well as recently solved NMR and X-ray structures spanning a higher size range. Each calculation was repeated considering different network definitions, as determined by the inflexion point (d_0) and the power term (p) in the distance function. Eigenvectors obtained for each protein and each combination of parameters (a point in the *(p, d)* map) defined an intrinsic deformation pattern, which was then compared with deformation patterns derived from; i) *MD ensembles*, ii) *NMR ensembles*, iii) conformational transitions in PDB. In the first case we center our analysis in the first 10 eigenvectors from both ENM and MD simulations computing the similarity index (γ_{10}). For NMR ensembles (always much smaller than the MD ones) we limit the comparison to just the first three eigenvectors (γ_3). Finally, in the case of experimentally know transitions, which are described by a single displacement vector, we determine the overlap between this vector and the 5 lowest-frequency modes eigenspace (γ_5), which usually covers most of the conformational change. The best-overlapped vector pair (γ_{max}) is also analyzed.

By changing in a systematic manner p and d_o we can scan virtually all the reasonable network definitions for a given structure (see *Figure 5. 9*), and check how the similarity between normal modes and a reference flexibility source changes with network topology. In general, the lowest similarities (around 0.6-0.7) are observed between ENMs and MD samplings (*Figure 5. 11*A, note the blue-black maps), followed by X-ray closed conformers (*Figure 5. 11*B) and ENM/NMR ensembles (values near 0.7 in some regions, *Figure 5. 11*C), while the highest similarities are obtained when ENMs eigenvectors are compared with X-ray experimental transitions from open states, with nearly homogeneous maps with values above 0.8 for all network structures (*Figure 5. 12*). These differences in maximal overlaps can be explained by the different timescale of the motions sampled by each one of the methods investigated. Clearly, longer time scale movements, as those involved in experimental transitions, are those better reproduced by ENMs, followed by NMR ensembles and MD simulations, which tend to sample more local motions.

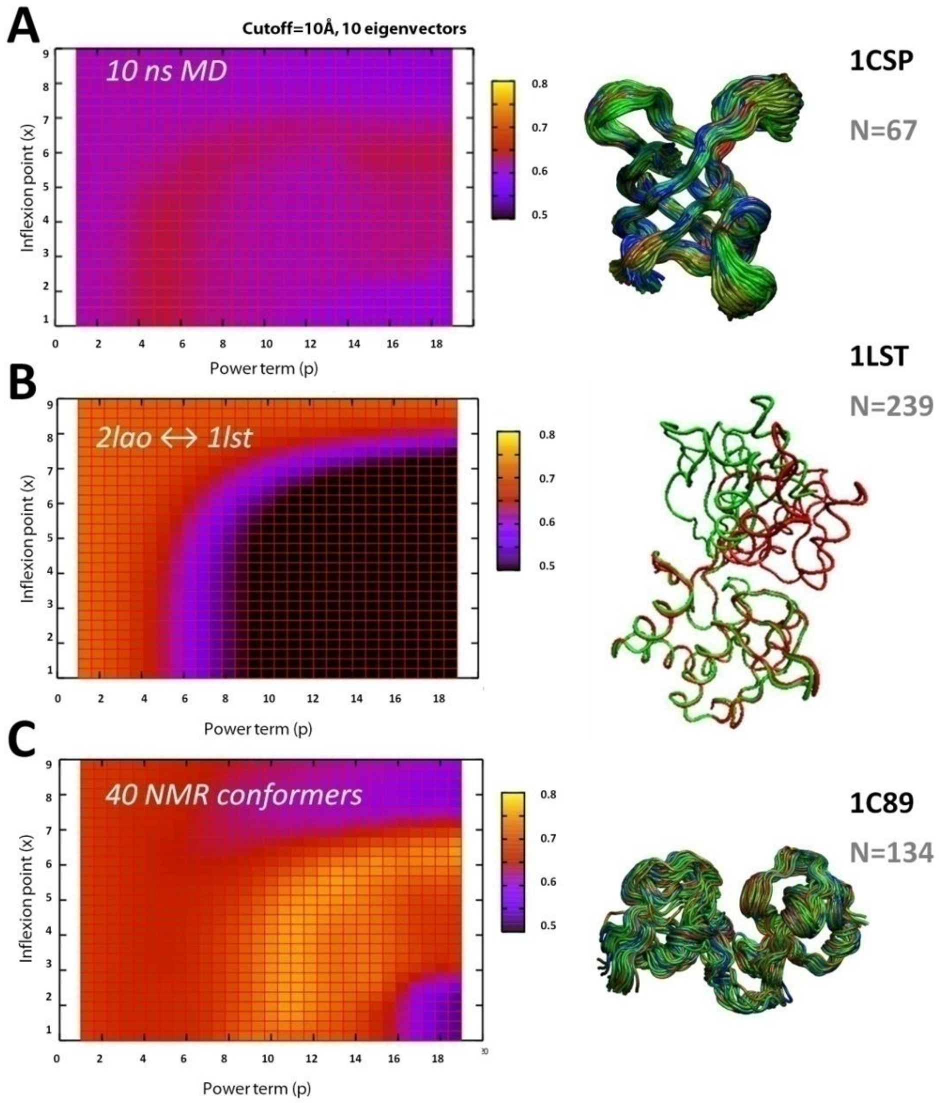


Figure 5. 11 **Dot product (p, x_0) space between NMA and different flexibility sources.**
Note the square-hyperbola transition between a region of low overlap and a region of high overlap. A) MD similarity for the model protein 1csp. B) X-ray bound/unbound pair. Dot Product between the first five normal modes of the open structure 1lstA and the vector driving the transition 1lst -> 2lao C) NMR ensemble similarity for the small protein 1c89. Code color: black=0.5 to yellow=0.8.

However, in spite of the different flexibility sources examined, and the different maximal similarities they reach, we found rather homogeneous gradients in *(p, d)* overlap maps. As can be seen in *Figure 5. 11* and *Figure 5. 12* square hyperbola-like boundaries between high and low overlap regions appear in the virtually all systems studied following a function of the kind:

$$d \propto p_0 + \frac{p}{1+p}$$ (Equation 5. 12 Hyperboloid boundary)

Where *d* is the optimal equilibrium distance, *p* is the power term at the boundary and p_0 its intersection with the axis at *d=0* (with values ranging typically from 4 to 10 depending on protein size). These boundaries mark transitions between high overlap (over 0.7- 0.8, in orange-yellow) to low overlap (below 0.5, in black) regions in the *(p, d_0)* space. In general, the differences in the maxima and minima found by systematic exploration of the parameter space are more relevant in the case of X-ray conformers, especially from the closed states (*Figure 5. 12*, 2nd column).

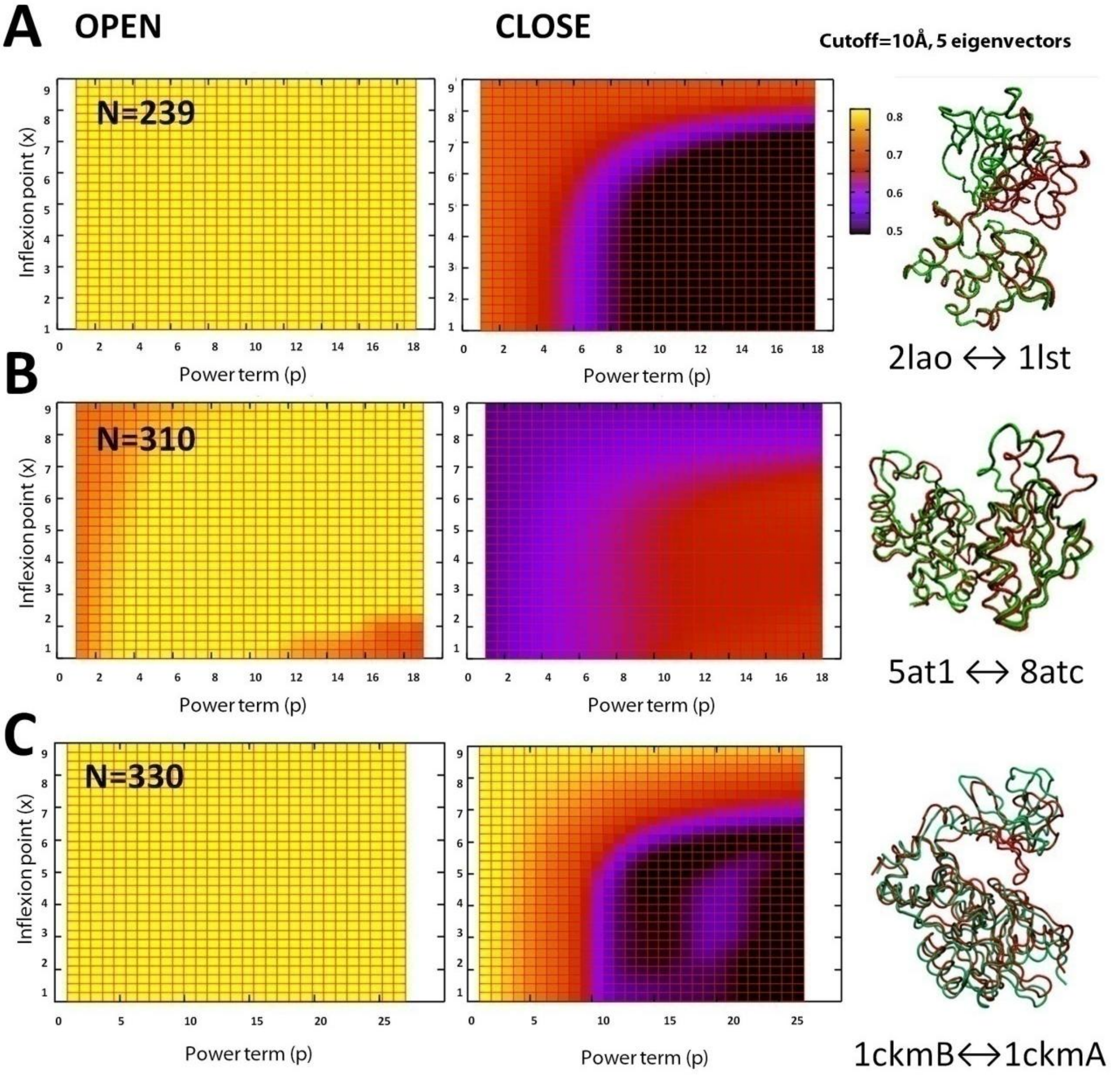


Figure 5. 12 Dot product (p, x_0) space between NMA and transitions from open or closed conformers.

Examples of extreme behaviors in networks for transitions from open (left) and closed (right) states (cutoff=10, 5 eigenvectors). a) 2lao to 1lst b) 5at1 to 8atc c) 1ckmB to 1ckmA

Then, we examined *if* and *how* the distribution of the similarity across the *(p, d)* maps is related to the intrinsic properties of the network. We found that most protein structures display analogous gradients for the examined topological properties when p and d change, and that therefore these properties are directly related with the ability of a network to trace biological motions, with correlations between the dot product *(p, d)* maps and topology *(p, d)* maps with absolute correlation coefficients often over 0.5. We found that, for most protein networks, the shortest paths oscillate between 0 and 6 links (R_c= 8) and between 0 and 3 (R_c=10), whereas the diameter oscillates between 0 and 10 (for R_c=8) and 0-5 (R_c=10). On the other hand, the clustering coefficient oscillates between 0.1 and 0.6 whereas assortativity ranges widely from 0.1 to 1. As we expected*,* the maxima and minima of these topological properties appear to be strongly related with the distribution of dot products and thus with the ability of the explored networks to trace protein flexibility.

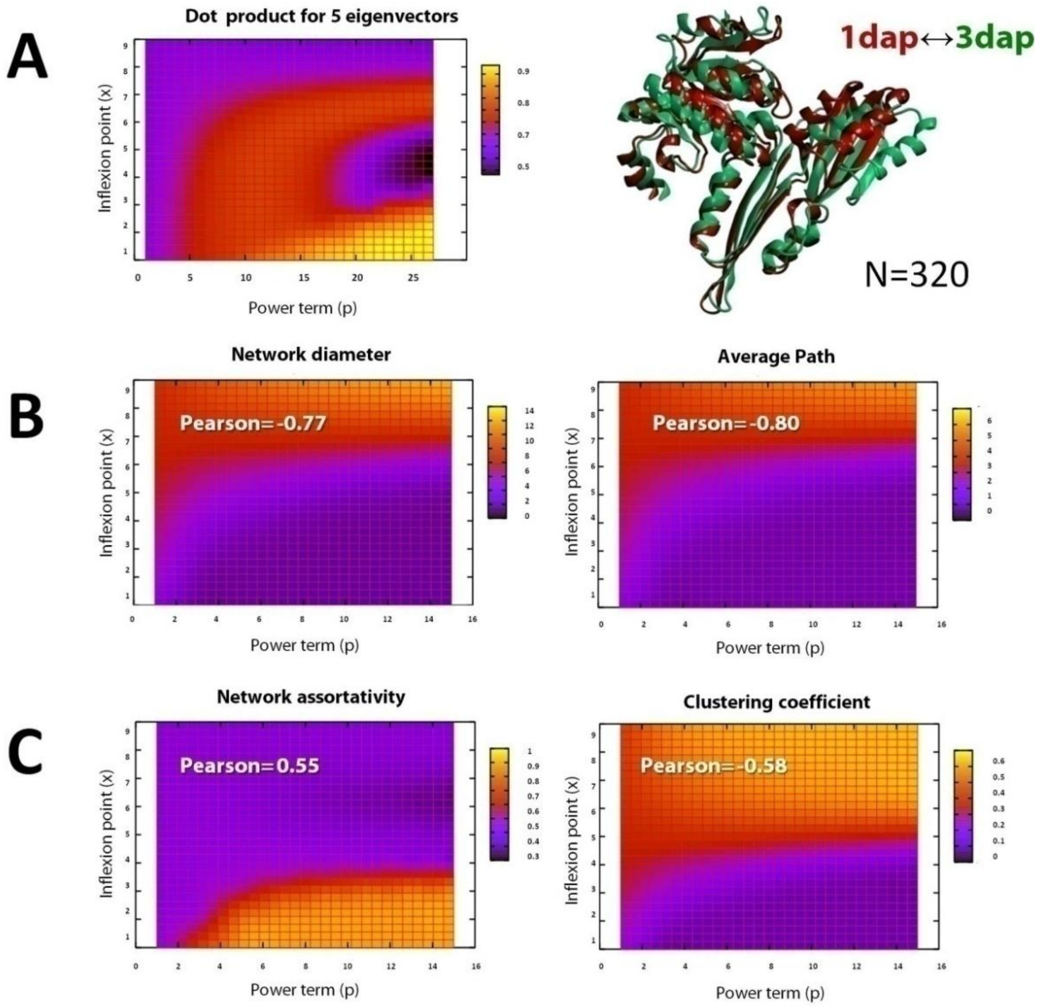


Figure 5. 13 **Comparison between dot product distribution and network properties for a representative case in a X-ray transition from the closed state (1dap).**

Note how highest overlaps (A) are related to small network diameters (2-6) and average paths (< 3) connecting the residues (B), and high assortativity values (0.9) and low clustering coefficients (<0.2).

In virtually all proteins examined, the network properties display very significant correlations with protein flexibility *(p, d)* maps. We found that in most cases, short average path lengths, small diameters and low clustering coefficients are significantly associated with dot product maxima, whereas assortativity coefficients display the opposite trend. For example, in a typical X-ray transition such as *1dap ↔3dap*, the lowest frequency space displays the above mentioned circular pattern with maximal overlaps at the lower-right corner (*Figure 5. 13, A)*. Inspection of the network properties *(p, d)* maps for network properties reveal similar hyperbola-like boundaries, with maxima or minima at either side. The overlap maxima display strong negative correlations with *cutoff-dependent* properties such as network diameter and average path (in the order of -0.8) (*Figure 5. 13, B)*. On the other side, *cutoff-independent* properties show sharp transitions at the lower right corner of the *(p, d)* maps, which also correlate with overlap maxima: whereas clustering coefficients also display a negative correlation (-0.58) with dot product maxima (*Figure 5. 13, C left)*, assortativity strongly correlates (+0.55) with the region of best overlap between ENMs and the conformational change (*Figure 5. 13, C right*). The best representation of protein flexibility is achieved in networks with assortativity coefficients near 1 and very low clustering coefficients near 0 (lower right corner in the corresponding maps), and a very compact topology characterized by small network sizes (2-6 links) and short paths (< 3 links). In a protein dynamics system, a short average path length and a small diameter indicates that residues are connected at a very local level. This suggests that a fast transmission of dynamical correlations across the C-alpha network is needed to capture experimental flexibility, suggesting also a greater sensitivity to allosteric regulation. On the other side, the requirement of low clustering coefficients, which measure the cohesiveness and density of contacts, indicates that although strongly connected at a local level, residue networks must be sparse enough to allow for a good description of protein motions. The assortativity coefficients close to 1 indicate the preference of highly connected residues (hubs) to connect to other hubs. Overall, this suggests that a highly selective connectivity between hubs (core residues at each structure element), but not high local cohesiveness (which could block structural motions), is required for the network to capture conformational changes. These results are in agreement with the findings exposed in **Chapter 3**, which suggest an important role for local connectivity in the description of functional motions.

5.5 Summary

We have seen in this Chapter several applications of the ED-ENM method to sample local motions and large-scale transitions. We have developed a method to trace physical pathways between different conformational states using a Langevin Dynamics implementation of the ED-ENM potentials biased by a Maxwell Demon. We also present an alternative implementation where ED-ENM modes are used to bias atomistic trajectories and explore more efficiently the conformational space. Then, we present a detailed study in which the hybrid Langevin/ENM method is used to analyze in depth the intrinsically encoded correlated motions in beta-sheets structures. We show that these local motions mediate twisting and bending of these structural elements which can be propagated across the backbone to trigger larger scale domain rearrangements. Finally, we further study the relations between the network properties of the backbone topology and the dynamics of proteins.

5.6 Publications from this chapter

Sfriso P., Emperador A., **Orellana L.**, Hospital A., Gelpí J.L., Orozco M. (2012) **Finding Conformational Transition Pathways from Discrete Molecular Dynamics Simulations.** *J. Chem. Theory Comput.* 8 (11): p 4707–4718

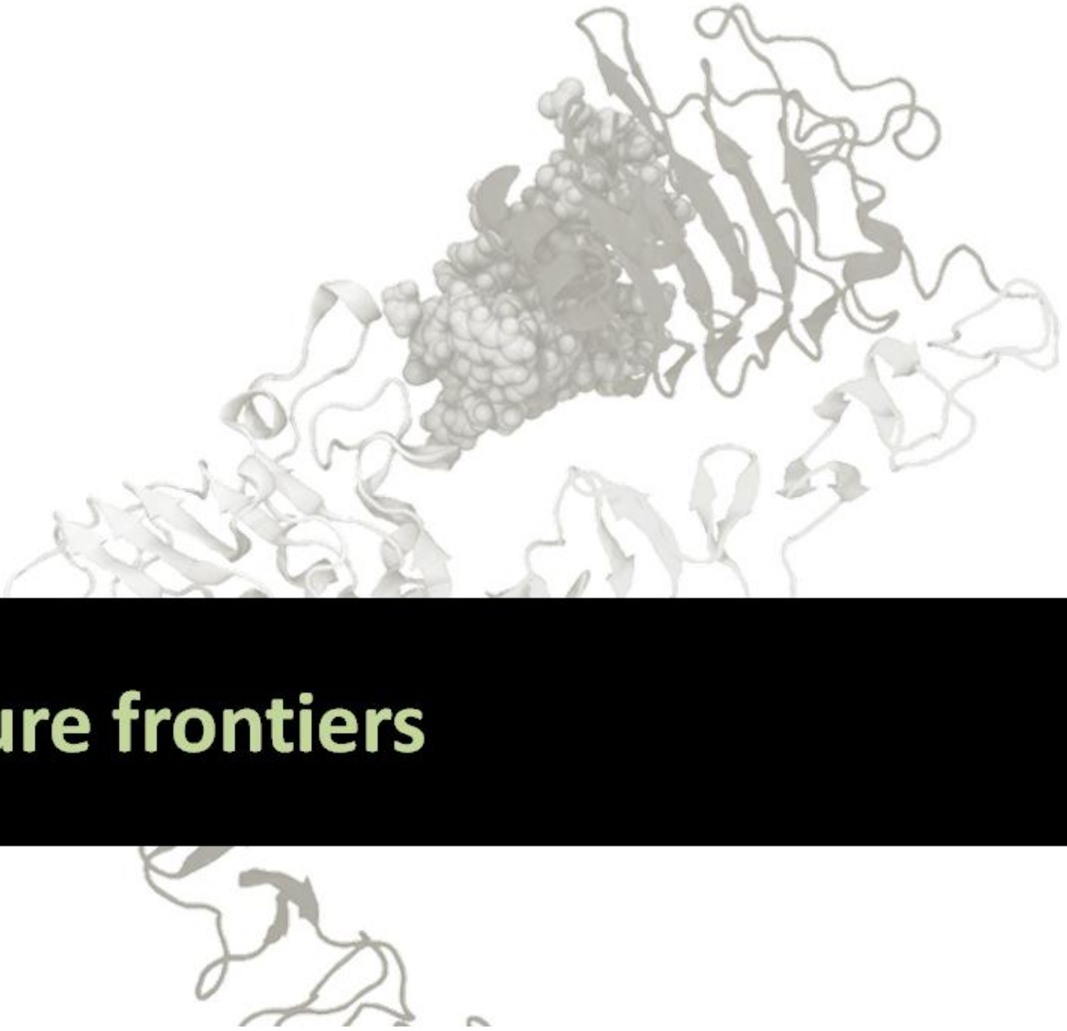
Fenwick R.B.*, **Orellana L.***, Esteban-Martin S., Salvatella X. and Orozco M. (2013) **Correlated Motions in β-sheets are an Inherent Property of Proteins.** *Nature Communications* (Second Revision) **Equal contribution*

Orellana L., and Orozco M. (2014) **Elastic Network-Brownian Dynamics transition pathways in large-scale conformational changes**. (*in preparation*)

Orellana L., López-Blanco J.R., Chacón P., Orozco M. (2014) **Exploring the link between residue network properties and protein dynamical behavior by fast dihedral NMA**. *(in preparation)*

“Computers are incredibly fast, accurate, and stupid: humans are incredibly slow, inaccurate and brilliant; together they are powerful beyond imagination”

Albert Einstein



6- Future frontiers

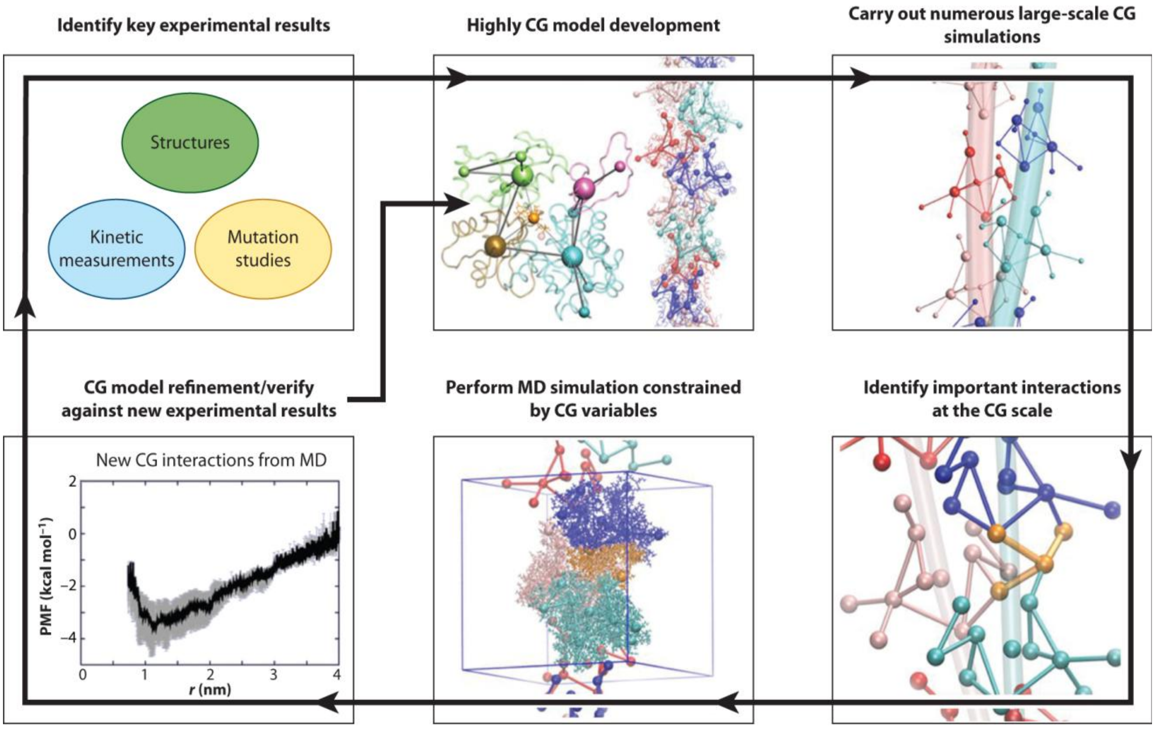
This page intentionally left blank

Chapter 6 Conclusions and future frontiers: From physics to medicine?

In this final part of the present thesis, we will summarize and discuss briefly the main results obtained and their relevance for the field of protein dynamics: from the interface with chemistry and physics (**Chapter 3**), the ultimate applications of theoretical predictions in biomedicine (**Chapter 4**) or the connections with complex networks theory (**Chapter 5**). A final personal view is given on the challenges in the simulation field for the upcoming years.

6.1 Atomistic versus coarse-grained simulations: sampling conformational landscapes.

As we saw in **Chapter 3**, the principal motions of a structure are strongly encoded in the backbone, nearest-neighbors covalent contacts. We have also seen that these shape-encoded motions are very similar irrespective of the source of flexibility: from theoretical methods such as ENM or ED, to empirical data from X-ray and NMR ensembles; as a rough estimation, one can say that all of them agree around 50-70% in the directions of the collective motions. The agreement between coarse-grained and atomistic, experimental and theoretical modes is very remarkable in proteins where large conformational rearrangements of structural elements occur. The agreement between MD and ENMs in the collective motions strongly demonstrates that coarse-graining is indeed an excellent alternative to guide atomistic simulations. Although to dissect the fine details of conformational transitions and protein mechanism atomistic simulations are clearly decisive, coarse-graining can help to focus on the interesting regions of the energy landscape. As a recent review by Gregory Voth envisions (Saunders & Voth, 2013), coarse-grained simulations will be the best-suited starting point to make sense of biological data and guide computationally demanding long-timescale MD simulations. A striking example of convergence of coarse-grained and atomistic predictions is the HER receptor, where near-microsecond long MD simulations sample almost perfectly the predicted normal modes. We have also seen in **Chapter 4** that the non-covalent, long range interactions (the out-diagonal contacts in the topology matrix) which connect the interfaces between distant structural elements also play a key modulatory function in the soft motions. In fact, a small perturbation in these contacts can shift dramatically the accessibility (i.e. the energies/frequencies) of the soft modes– making easier for example, to explore rare configurations, or a wider region in the conformational landscape, in faster timescales.


Identify key experimental results
Highly CG model development
Carry out numerous large-scale CG simulations
Structures
Kinetic measurements
Mutation studies
CG model refinement/verify against new experimental results
Perform MD simulation constrained by CG variables
Identify important interactions at the CG scale
New CG interactions from MD
PMF (kcal mol⁻¹)
2
0
−2
−4
0
1
2
3
4
r (nm)

.

Figure 6. 1 **A paradigm for multiscale problems.**
Combining coarse-grained and atomistic simulations to understand biomolecular phenomena and connect simulations and experiments. The development of powerful coarse-grained approaches can help to focus the computationally-expensive atomistic simulations and connect data from different scales to unveil complex biological mechanisms.

Published in: Marisa G. Saunders; Gregory A. Voth; *Annu.Rev.Biophys.* 2013, 42, 73-93. 10.1146/annurev-biophys-083012-130348

6.2 Disease mutations: a key to uncover protein mechanisms?

As we saw in **Chapter 4**, the natural, collective motions of a protein are strongly encoded in its overall shape. However, we have also seen that mutations hitting at sensitive points, such as interdomain hinges and interfaces, can drive dramatic changes in the near-equilibrium dynamics. What’s the meaning of such weakness in control proteins, which play highly delicate roles in the cell, such as the EGF-receptor? Multidomain sensor proteins must clearly signal cell events that require a complex modulation of which functional surfaces are going to be exposed at every moment. Precisely, these sensitive regions are often the target of environmental cues such as ligands, whose binding can drive conformational changes. Mutations targeting control sites may therefore shift the equilibrium between different structural states, which may result in protein malfunction or deregulation. In other words, mutations at sensitive regions can perturb the sampling of the available conformational space. If a particular structure has evolved to perform a given transition upon ligand binding, a mutation can favor the same shape-encoded movements in the absence of triggering signals. The sensitivity of such structures, required for signaling, becomes a weakness exploited by cancer cells, which undergo fast microevolution processes which favor those cells harboring growth and survival promoting mutations. This is particularly relevant in the case of multidomain proteins. As we have seen in the case of the HER1 receptor, a simple point mutation targeting non-covalent bonds in a critical interface can accelerate the intrinsic dynamics allowing a structure to sample the conformational space in a much faster timescale, revealing transitions that can elude even microsecond-long simulations. This finding brings the following question: *Could be use mutations as landmarks to sample more efficiently the conformational space?* Instead of performing blindly long microsecond simulations, a faster way to explore exhaustively the possible configurations of a given structure could be to simulate mutations that increase the local flexibility at critical points (interdomain/intermodular hinges and surfaces), in order to highlight the less visited regions of the conformational landscape.

6.3 Summary and conclusions of this thesis

As a final summary, the main findings of this thesis are the following:

1. ***Collective dynamics can be described by coarse-grained methods, and specially, by ENMs, which give results of an accuracy comparable to MD regarding large-scale flexibility***
2. ***As ENMs demonstrate, protein collective dynamics is shape-encoded in the nearest-neighbor covalent connectivity of the backbone that dictates the fold; in other words, it depends on the arrangement of large structural elements***
3. ***Long-range contacts play a critical role governing the accessibility of the different large-scale motions available for a particular structure***
4. ***These critical contacts can target the surfaces between different structural elements, and upon mutation, shift dramatically the essential movements; since they connect the different structural elements, they can block or allow different motions and thus restrict the flexibility space***

Although protein dynamics is a challenging problem, most available methods, experimental and theoretical, seem to agree in one point: that the large-scale motions of a structure are largely shape and topology dependent. This unity of results demonstrates that research is evolving in the correct direction, and that discrepancies are not as relevant as it may seem. Experimental and theoretical approaches are telling us that the landscape of motions available for a structure is clearly limited and dependent on his shape. The challenge is to distinguish which are the physical ones, and among them, those selected by evolution, which can also play sometimes with randomness. Coarse-grained approaches should not be considered a minor approach compared to atomistic simulations; on the contrary, are well-suited to undertake problems in the length scale where conformational changes occur. However, atomistic simulations are undeniably needed to sort out the detailed mechanisms and the energies involved in these processes.

6.4 A personal view: bridging the gap between scales

The Award of the Nobel Prize in Chemistry this year to the founders of the biomolecular simulation, **Martin Karplus**, **Michael Levitt** and **Arieh Warshel**, represents an important and necessary recognition to our field from the research community. In 1967, Levitt and Warshel first designed a computer program that used traditional, Newtonian physics to perform energy minimization of protein structures, the *CFF (Consistent Force-Field)*, which was later used by Karplus and coworkers as a launching point for the historical first molecular dynamics simulation of a protein, published ten years later (See a nice review in (Levitt, 2001)). Since the bare 3 picoseconds of BPTI simulated by McCammon, Gelin and Karplus (J A McCammon et al., 1977) to the millisecond dynamics by the Anton Supercomputer (Shaw et al., 2010), structural bioinformatics has come a very long way. Although being firmly rooted in rigorous physical laws – or just because of that – simulations were constrained to very small timescales of apparently little biological significance, and therefore have not been considered as scientific tools with the same category as *"real"* (wet-lab, physical or material) experiments in Life sciences; computations were for many years considered *"second-class"* evidence by most experimentalists. Now, as we approach millisecond scales and start to be able to deal with huge systems, the situation is changing and simulations start to occupy an honorable place in Chemistry and Life Sciences.

I think Science should be a table with three interconnected legs: theory, simulation and experimentation. Simulations are indeed the most rigorous bridge – a quantitative one - between an experiment and any theory behind. However, as the order of magnitude and the complexity of a system increase, it is true that more and more difficult becomes the modeling and validation by simulations. For this reason, biology has escaped so far from the rigorous approach of physics and chemistry, and actually has done quite well using mostly qualitative models – the pervasive boxes and arrows, "cause-effect" schemes that almost seem to substitute the real things everywhere. As the polemic and funny paper by Yuri Lazebnik in *Cancer Cell* criticized, ***"Can a biologist***

fix a radio?” (Lazebnik, 2002), most biologists elude the hard-science quantitative approach because of an irrational fear of Mathematics, and are happy accumulating massive data with little rational behind:

“I started to contemplate how biologists would determine why my radio does not work and how they would attempt to repair it….Eventually, all components will be cataloged, connections between them will be described, and the consequences of removing each component or their combinations will be documented. This will be the time when the question, previously obscured by the excitement of productive research, would have to be asked: Can the information that we accumulated help us to repair the radio? “

During the last century, the development of molecular biology and structural techniques has provided a tremendous quantity of data that requires a rational framework and at the same time, will allow us, for the first time, to unveil the links that connect observations from very different scales of Life: from the molecular level, to the cell, tissue and organism scale. Up to date, simulations usually served only as a complement to experiments in the field of structural biology. However, as the increasing computational power is allowing us to reach the biologically relevant scales, the situation will revert, and simulations can start to suggest novel hypothesis and guide experiments. By pushing the limits of simulation techniques, we can begin to envision and construct rational and quantitative models from the growing amount of biological data. From structural bioinformatics to systems biology, probably in the next and exciting years, we will be able to dare, for the first time, to explain how the human body works and how the normal function is perturbed by disease down to the molecular detail. And perhaps we can start to repair radios…

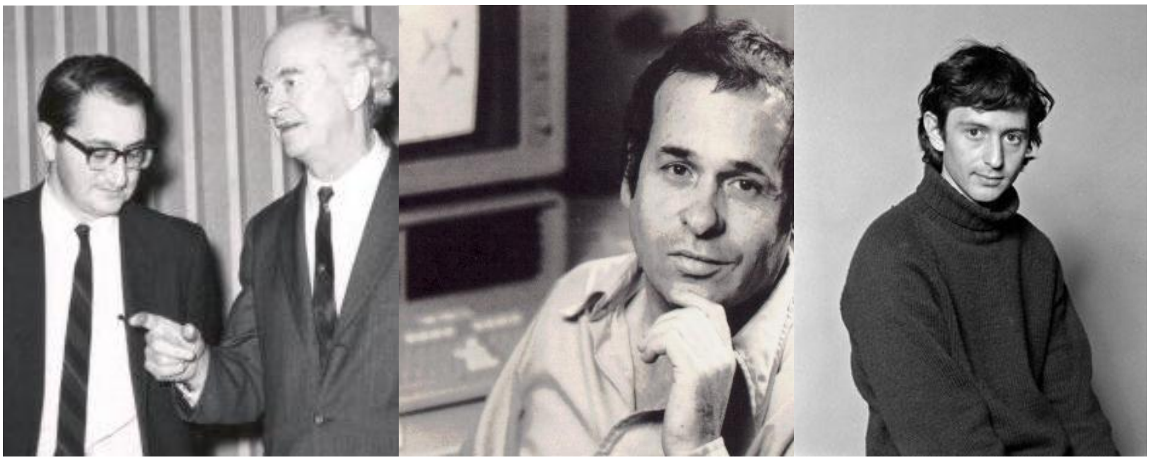


Figure 6. 2 **The Nobel Prize in Chemistry 2013.**
The Nobel committee this year has recognized the founding fathers of computational structural biology back in the 70s. *Martin Karplus with his PhD advisor Linus Pauling (left) Arieh Warshel (center) and Michael Levitt (right).*

This page intentionally left blank

References

Abseher, R., Horstink, L., Hilbers, C. W., & Nilges, M. (1998). Essential spaces defined by NMR structure ensembles and molecular dynamics simulation show significant overlap. *Proteins*, *31*(4), 370–82.

Acuner Ozbabacan, S. E., Gursoy, A., Keskin, O., & Nussinov, R. (2010). Conformational ensembles, signal transduction and residue hot spots: application to drug discovery. *Current opinion in drug discovery & development*, *13*(5), 527–37.

Alder, B. J., & Wainbright, T. E. (1957). Phase Transition for a Hard Sphere System. *Journal of Chemical Physics*, *27*, 1208.

Alexandrov, V., Lehnert, U., Echols, N., Milburn, D., Engelman, D., & Gerstein, M. (2005). Normal modes for predicting protein motions : A comprehensive database assessment and associated Web tool, 633–643. doi:10.1110/ps.04882105.amplitude

Alvarado, D., Klein, D. E., & Lemmon, M. A. (2009). ErbB2 resembles an autoinhibited invertebrate epidermal growth factor receptor. *Nature*, *461*(7261), 287–291. doi:10.1038/nature08297

Alvarado, D., Klein, D. E., & Lemmon, M. A. (2010). Structural Basis for Negative Cooperativity in Growth Factor Binding to an EGF Receptor. *Cell*, *142*(4), 568–579. doi:10.1016/j.cell.2010.07.015

Amadei, A., Linssen, A. B., & Berendsen, H. J. (1993). Essential dynamics of proteins. *Proteins*, *17*(4), 412–425.

Amadei, A., Linssen, A. B., De Groot, B. L., Van Aalten, D. M., & Berendsen, H. J. (1996). An efficient method for sampling the essential subspace of proteins. *Journal of biomolecular structure dynamics*, *13*(4), 615–625.

Antal, M. A., Bode, C., & Csermely, P. (2009). Perturbation waves in proteins and protein networks: Applications of percolation and game theories in signaling and drug design. *Current Protein and Peptide Science*, *10*(2), 161–172.

Arenas, A., Fernandez, A., & Gomez, S. (2007). Analysis of the structure of complex networks at different resolution levels. *New Journal of Physics*, *10*(5), 23.

Arkhipov, A., Shan, Y., Das, R., Endres, N. F., Eastwood, M. P., Wemmer, D. E., … Shaw, D. E. (2013). Architecture and membrane interactions of the EGF receptor. *Cell*, *152*(3), 557–69. doi:10.1016/j.cell.2012.12.030

Arold, S. T., Hoellerer, M. K., & Noble, M. E. M. (2002). The structural basis of localization and signaling by the focal adhesion targeting domain. *Structure (London, England : 1993)*, *10*(3), 319–327. doi:10.1016/S0969-2126(02)00717-7

Atilgan, A. R., Durell, S. R., Jernigan, R. L., Demirel, M. C., Keskin, O., & Bahar, I. (2001). Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophysical journal*, *80*(1), 505–15. doi:10.1016/S0006-3495(01)76033-X

Babu, M. M., Van Der Lee, R., De Groot, N. S., & Gsponer, J. (2011). Intrinsically disordered proteins: regulation and disease. *Current Opinion in Structural Biology*, *21*(3), 432–440.

Bagler, G, & Sinha, S. (2005). Network properties of protein structures. *Physica A: Statistical Mechanics and its Applications*, *346*(1-2), 27–33. doi:10.1016/j.physa.2004.08.046

Bagler, Ganesh, & Sinha, S. (2007). Assortative mixing in Protein Contact Networks and protein folding kinetics. *Bioinformatics (Oxford, England)*, *23*(14), 1760–7. doi:10.1093/bioinformatics/btm257

Bahar, I, Atilgan, A. R., & Erman, B. (1997). Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Folding design*, *2*(3), 173–181.

Bahar, Ivet, & Rader, A. J. (2005). Coarse-grained normal mode analysis in structural biology. *Current Opinion in Structural Biology*, *15*(5), 586–592.

Barrat, A., Barthélemy, M., Pastor-Satorras, R., & Vespignani, A. (2004). The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(11), 3747–3752.

Baselga, J. (2002). Why the epidermal growth factor receptor? The rationale for cancer therapy. *The oncologist*, *7 Suppl 4*(suppl 4), 2–8.

Beckstein, O., Denning, E. J., Perilla, J. R., & Woolf, T. B. (2009). Zipping and Unzipping of Adenylate Kinase: Atomistic Insights into the Ensemble of Open&#xa0;↔&#xa0;Closed Transitions. *Journal of Molecular Biology*, *394*(1), 160–176. doi:http://dx.doi.org/10.1016/j.jmb.2009.09.009

Bejan, A., & Lorente, S. (2010). The constructal law of design and evolution in nature. *Philosophical Transactions of the Royal Society of London - Series B: Biological Sciences*, *365*(1545), 1335–1347.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., … Bourne, P. E. (2000). The Protein Data Bank. (F. H. Allen, G. Berghoff, & R. Sievers, Eds.)*Nucleic Acids Research*, *28*(1), 235–242.

Bernadó, P., Mylonas, E., Petoukhov, M. V, Blackledge, M., & Svergun, D. I. (2007). Structural Characterization of Flexible Proteins Using Small-Angle X-ray Scattering. *Journal of the American Chemical Society*, *129*(17), 5656–5664. doi:10.1021/ja069124n

Blake, C. C. F., Koenig, D. F., Mair, G. A., North, A. C. T., Phillips, D. C., & Sarma, V. R. (1965). Structure of hen egg-white lysozyme, a three dimensional fourier synthesis at 2~Ångstroms resolution. *Nature*, *206*(4986), 757–761.

Bloch, F., Hansen, W., & Packard, M. (1946). The Nuclear Induction Experiment. *Physical Review*, *70*(7-8), 474–485. doi:10.1103/PhysRev.70.474

Braxenthaler, M., Unger, R., Auerbach, D., Given, J. A., & Moult, J. (1997). Chaos in protein dynamics. *Proteins*, *29*(4), 417–425. doi:10.1002/(SICI)1097-0134(199712)29:4<417::AID-PROT2>3.0.CO;2-5 [pii]

Brunger, A. T., Strop, P., Vrljic, M., Chu, S., & Weninger, K. R. (2011). Three-dimensional molecular modeling with single molecule FRET. *Journal of Structural Biology*, *173*(3), 497–505. doi:http://dx.doi.org/10.1016/j.jsb.2010.09.004

Brüschweiler, R. (1995). Collective protein dynamics and nuclear spin relaxation. *The Journal of Chemical Physics*, *102*(8).

Camps, J., Carrillo, O., Emperador, A., Orellana, L., Hospital, A., Rueda, M., … Orozco, M. (2009). FlexServ: an integrated tool for the analysis of protein flexibility. *Bioinformatics (Oxford, England)*, *25*(13), 1709–10. doi:10.1093/bioinformatics/btp304

Carrillo, O., Laughton, C. A., & Orozco, M. (2012). Fast Atomistic Molecular Dynamics Simulations from Essential Dynamics Samplings. *Journal of Chemical Theory and Computation*, *8*(3), 792–799. doi:10.1021/ct2007296

Case, D. A. (1994). Normal mode analysis of protein dynamics. *Current Opinion in Structural Biology*, *4*(2), 285–290. doi:10.1016/S0959-440X(94)90321-2

Chao, G., Cochran, J. R., & Wittrup, K. D. (2004). Fine Epitope Mapping of anti-Epidermal Growth Factor Receptor Antibodies Through Random Mutagenesis and Yeast Surface Display. *Journal of Molecular Biology*, *342*(2), 539–550. doi:10.1016/j.jmb.2004.07.053

Chennubhotla, C., & Bahar, I. (2007). Markov methods for hierarchical coarse-graining of large protein dynamics. *Journal of computational biology*, *14*(6), 765–76. doi:10.1089/cmb.2007.R015

Choe, S., & Sun, S. X. (2007). Bending elasticity of anti-parallel beta-sheets. *Biophysical journal*, *92*(4), 1204–1214.

Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., … Kollman, P. A. (1995). A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of the American Chemical Society*, *117*(19), 5179–5197. doi:10.1021/ja00124a002

Court, T. P. (2009). Flexible ligand docking to multiple receptor conformations: a practical alternative, *18*(2), 178–184.

Csermely, P., Palotai, R., & Nussinov, R. (2010). Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. *Trends in Biochemical Sciences*, *35*(10), 539–546.

Davis, I. W., Arendall, W. B., Richardson, D. C., & Richardson, J. S. (2006). The backrub motion: how protein backbone shrugs when a sidechain dances. *Structure (London, England : 1993)*, *14*(2), 265–274. doi:10.1016/j.str.2005.10.007

Dawson, J. P., Bu, Z., & Lemmon, M. a. (2007). Ligand-induced structural transitions in ErbB receptor extracellular domains. *Structure*, *15*(8), 942–54. doi:10.1016/j.str.2007.06.013

De Groot, B. L., Amadei, A., Scheek, R. M., Van Nuland, N. A., & Berendsen, H. J. (1996). An extended sampling of the configurational space of HPr from E. coli. *Proteins*, *26*(3), 314–322.

De Groot, B. L., Amadei, A., Van Aalten, D. M. F., & Berendsen, H. J. C. (1996). Towards an exhaustive sampling of the configurational spaces of the two forms of the peptide hormone guanylin. *J Biomol Str Dyn*, *13*(5), 741–751.

De Groot, B. L., van Aalten, D. M., Amadei, A., & Berendsen, H. J. (1996). The consistency of large concerted motions in proteins in molecular dynamics simulations. *Biophysical journal*, *71*(4), 1707–1713.

Del Sol, A., Fujihashi, H., Amoros, D., & Nussinov, R. (2006). Residues crucial for maintaining short paths in network communication mediate signaling in proteins. *Molecular systems biology*, *2*, 2006.0019. doi:10.1038/msb4100063

Del Sol, A., Tsai, C.-J., Ma, B., & Nussinov, R. (2009). The origin of allosteric functional modulation: multiple pre-existing pathways. *Structure (London, England : 1993)*, *17*(8), 1042–50. doi:10.1016/j.str.2009.06.008

Di Paola, L., De Ruvo, M., Paci, P., Santoni, D., & Giuliani, A. (2012). Protein Contact Networks: An Emerging Paradigm in Chemistry. *Chemical Reviews*, *113*(3), 1598–1613. doi:10.1021/cr3002356

Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische Mathematik*, *1*(1), 269–271. doi:10.1007/BF01386390

Dobbins, S. E., Lesk, V. I., & Sternberg, M. J. E. (2008). Insights into protein flexibility: The relationship between normal modes and conformational change upon protein-protein docking. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(30), 10390–5. doi:10.1073/pnas.0802496105

Dokholyan, N. V, Li, L., Ding, F., & Shakhnovich, E. I. (2002). Topological determinants of protein folding. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(13), 8637–41. doi:10.1073/pnas.122076099

Dos Santos, H. G., Klett, J., Méndez, R., & Bastolla, U. (2013). Characterizing conformation changes in proteins through the torsional elastic response. *Biochimica et biophysica acta*, *1834*(5), 836–46. doi:10.1016/j.bbapap.2013.02.010

Du, Y., Yang, H., Xu, Y., Cang, X., Luo, C., Mao, Y., … Jiang, H. (2012). Conformational Transition and Energy Landscape of ErbB4 Activated by Neuregulin1β: One Microsecond Molecular Dynamics Simulations. *Journal of the American Chemical Society*, *134*(15), 6720–31. doi:10.1021/ja211941d

DuBay, K. H., Geissler, P. L., & Bothma, J. P. (2011). Long-Range Intra-Protein Communication Can Be Transmitted by Correlated Side-Chain Fluctuations Alone. *PLoS Computational Biology*. doi:10.1371/journal.pcbi.1002168

Durand, P., Trinquier, G., & Sanejouand, Y.-H. (1994). A new approach for determining low-frequency normal modes in macromolecules. *Biopolymers*, *34*(6), 759–771. doi:10.1002/bip.360340608

Elofsson, A., & Nilsson, L. (1993). How consistent are molecular dynamics simulations? Comparing structure and dynamics in reduced and oxidized Escherichia coli thioredoxin. *Journal of molecular biology*, *233*(4), 766–780.

Emberly, E. G., Mukhopadhyay, R., Tang, C., & Wingreen, N. S. (2004). Flexibility of beta-sheets: principal component analysis of database protein structures. *Proteins*, *55*(1), 91–8. doi:10.1002/prot.10618

Emperador, A., Carrillo, O., Rueda, M., & Orozco, M. (2008). Exploring the suitability of coarse-grained techniques for the representation of protein dynamics. *Biophysical Journal*, *95*(5), 2127–2138.

Eyal, E., Chennubhotla, C., Yang, L.-W., & Bahar, I. (2007). Anisotropic fluctuations of amino acids in protein structures: insights from X-ray crystallography and elastic network models. *Bioinformatics*, *23*(13), i175–i184.

Eyal, E., Yang, L.-W., & Bahar, I. (2006). Anisotropic network model: systematic evaluation and a new web interface. *Bioinformatics*, *22*(21), 2619–2627.

Fenwick, R. B., Esteban-Martín, S., Richter, B., Lee, D., Walter, K. F. a, Milovanovic, D., … Salvatella, X. (2011). Weak long-range correlated motions in a surface patch of ubiquitin involved in molecular recognition. *Journal of the American Chemical Society*, *133*(27), 10336–9. doi:10.1021/ja200461n

Fenwick, R. B., Orellana, L., Esteban-Martín, S., Orozco, M., & Salvatella, X. (2013). Correlated motions are a fundamental property of beta-sheets. *(submitted)*.

Ferguson, K. M. (2009). A structure-based view of Epidermal Growth Factor Receptor regulation, (215), 353–373. doi:10.1146/annurev.biophys.37.032807.125829.A

Ferguson, K. M., Berger, M. B., Mendrola, J. M., Cho, H.-S., Leahy, D. J., & Lemmon, M. a. (2003). EGF Activates Its Receptor by Removing Interactions that Autoinhibit Ectodomain Dimerization. *Molecular Cell*, *11*(2), 507–517. doi:10.1016/S1097-2765(03)00047-9

Fischer, E. (1894). Einfluss der Configuration auf die Wirkung der Enzyme. *Berichte der deutschen chemischen Gesellschaft*, *27*(3), 2985–2993. doi:10.1002/cber.18940270364

Flory, P. J., Gordon, M., & McCrum, N. G. (1976). Statistical Thermodynamics of Random Networks [and Discussion]. *Proceedings of the Royal Society A Mathematical Physical and Engineering Sciences*, *351*(1666), 351–380. doi:10.1098/rspa.1976.0146

Fraser, J. S., Clarkson, M. W., Degnan, S. C., Erion, R., Kern, D., & Alber, T. (2009). Hidden alternative structures of proline isomerase essential for catalysis. *Nature*, *462*(7273), 669–673.

Gabel, F., Bicout, D., Lehnert, U., Tehei, M., Weik, M., & Zaccai, G. (2002). Protein dynamics studied by neutron scattering. *Quarterly Reviews of Biophysics*, *35*(4), 327–367. doi:10.1017/S0033583502003840

Gan, H. K., Burgess, A. W., Clayton, A. H. A., Res, C., Onlinefirst, P., & Scott, A. M. (2012). Targeting of a Conformationally Exposed , Tumor-Specific Epitope of EGFR as a Strategy for Cancer Therapy. *Cancer Research*, *72*(12), 2924–30. doi:10.1158/0008-5472.CAN-11-3898

Gan, H. K., Walker, F., Burgess, A. W., Rigopoulos, A., Scott, A. M., & Johns, T. G. (2007). The epidermal growth factor receptor (EGFR) tyrosine kinase inhibitor AG1478 increases the formation of inactive untethered EGFR dimers. Implications for combination therapy with monoclonal antibody 806. *The Journal of biological chemistry*, *282*(5), 2840–50. doi:10.1074/jbc.M605136200

Garrett, T. P. J., Burgess, A. W., Gan, H. K., Luwor, R. B., Cartwright, G., Walker, F., … Johns, T. G. (2009). Antibodies specifically targeting a locally misfolded region of tumor associated EGFR. *Proceedings of the National Academy of Sciences*, *106*, 5082–5087.

Garrett, T. P. J., Mckern, N. M., Lou, M., Elleman, T. C., Adams, T. E., Lovrecz, G. O., … Parade, R. (2008). The Crystal Structure of a Truncated ErbB2 Ectodomain Reveals an Active Conformation , Poised to Interact with Other ErbB Receptors. *Structure*, *11*, 495–505.

Gerstein, M., & Krebs, W. (1998). A database of macromolecular motions. *Nucleic acids research*, *26*(18), 4280–90.

Go, N. (1990). A theorem on amplitudes of thermal atomic fluctuations in large molecules assuming specific conformations calculated by normal mode analysis. *Biophysical Chemistry*, *35*(1), 105–112.

Go, Nobuhiro, Noguti, T., & Nishikawa, T. (1983). Dynamics of a small globular protein in terms of low-frequency vibrational modes. *Biophysics*, *80*(June), 3696–3700. doi:10.1021/jp111420q

Gomez, S., Jensen, P., & Arenas, A. (2008). Analysis of community structure in networks of correlated data. *Physical Review E*, *80*(1), 5.

Grubmüller, H. (1995). Predicting slow structural transitions in macromolecular systems: Conformational flooding. *Physical Review E*, *52*(3), 2893–2906. doi:10.1103/PhysRevE.52.2893

Gunasekaran, K., Ma, B., & Nussinov, R. (2004). Is allostery an intrinsic property of all dynamic proteins? *Proteins*, *57*(3), 433–43. doi:10.1002/prot.20232

Guvench, O., & MacKerell, A. D. (2008). Comparison of protein force fields for molecular dynamics simulations. (A. Kukol, Ed.)*Methods In Molecular Biology Clifton Nj*, *443*, 63–88.

Ha-Duong, T., Basdevant, N., & Borgis, D. (2009). A polarizable coarse-grained water model for coarse-grained proteins simulations. *Chemical Physics Letters*, *468*(1-3), 79–82. doi:10.1016/j.cplett.2008.11.092

Hajdu, J., Neutze, R., Sjogren, T., Edman, K., Szoke, A., Wilmouth, R. C., & Wilmot, C. M. (2000). Analyzing protein functions in four dimensions. *Nat Struct Mol Biol*, *7*(11), 1006–1012.

Haliloglu, T., Bahar, I., & Erman, B. (1997). Gaussian Dynamics of Folded Proteins. *Physical Review Letters*, *79*(16), 3090–3093. doi:10.1103/PhysRevLett.79.3090

Halle, B. (2002). Flexibility and packing in proteins. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(3), 1274–9. doi:10.1073/pnas.032522499

Hamacher, K., & McCammon, J. A. (2006). Computing the Amino Acid Specificity of Fluctuations in Biomolecular Systems. *Journal of Chemical Theory and Computation*, *2*(3), 873–878. doi:10.1021/ct050247s

Hansmann, U. H., & Okamoto, Y. (1999). New Monte Carlo algorithms for protein folding. *Current opinion in structural biology*, *9*(2), 177–183. doi:10.1016/S0959-440X(99)80025-6

Henzler-Wildman, K., & Kern, D. (2007). Dynamic personalities of proteins. *Nature*, *450*(7172), 964–72. doi:10.1038/nature06522

Herzberg, G. (1945). *Molecular Spectra and Molecular Structure*. Princeton, New Jersey: D. Van Nostrand Company, Inc.

Hess, B. (2000). Similarities between principal components of protein dynamics and random diffusion. *Physical review. E*, *62*(6 Pt B), 8438–48.

Hess, B., Kutzner, C., Van Der Spoel, D., & Lindahl, E. (2008). GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *Journal of Chemical Theory and Computation*, *4*(3), 435–447. doi:10.1021/ct700301q

Heyduk, T. (2002). Measuring protein conformational changes by FRET/LRET. *Current Opinion in Biotechnology*, *13*(4), 292–296. doi:http://dx.doi.org/10.1016/S0958-1669(02)00332-4

Hinsen, K. (1998). Analysis of domain motions by approximate normal mode calculations. *Proteins*, *33*(3), 417–29.

Hinsen, Konrad. (2008). Structural flexibility in proteins: impact of the crystal environment. *Bioinformatics (Oxford, England)*, *24*(4), 521–8. doi:10.1093/bioinformatics/btm625

Hinsen, Konrad, Petrescu, A.-J., Dellerue, S., Bellissent-Funel, M.-C., & Kneller, G. R. (2000). Harmonicity in slow protein dynamics. *Chemical Physics*, *261*(1-2), 25–37. doi:10.1016/S0301-0104(00)00222-6

Horvath, G., Farkas, E., Boncz, I., Blaho, M., & Kriska, G. (2012). Cavemen Were Better at Depicting Quadruped Walking than Modern Artists: Erroneous Walking Illustrations in the Fine Arts from Prehistory to Today. *PLoS ONE*, *7*(12). doi:10.1371/journal.pone.0049786

Huang, P. H., Xu, A. M., & White, F. M. (2009). Oncogenic EGFR Signaling Networks in Glioma. *Science Signaling*, *2*(87), re6.

Istomin, A. Y., Gromiha, M. M., Vorov, O. K., Jacobs, D. J., & Livesay, D. R. (2008). New insight into long-range nonadditivity within protein double-mutant cycles. *Proteins*, *70*(3), 915–924. doi:10.1002/prot.21620

Jacobs, D. J., Rader, A. J., Kuhn, L. A., & Thorpe, M. F. (2001). Protein flexibility predictions using graph theory. *Proteins*, *44*(2), 150–165.

Jammalamadaka, S. R., & Sengupta, A. (2001). *Topics in circular statistics*. *Statistics* (Vol. 6, p. e20505). World Scientific Press. doi:10.1111/j.1442-2050.2010.01169.x

Jeong, J. I., Jang, Y., & Kim, M. K. (2006). A connection rule for alpha-carbon coarse-grained elastic network models using chemical bond information. *Journal of molecular graphics modelling*, *24*(4), 296–306.

Jernigan, R. L. (2007). Comparison of Experimental and Computed Protein Anisotropic Temperature Factors Department of Biochemistry , Biophysics and Molecular Biology. *Methods*, 89–96.

Jin, Z., Jin, L., Peterson, D. L., & Lawson, C. L. (1999). Model for lentivirus capsid core assembly based on crystal dimers of EIAV p26. *Journal of molecular biology*, *286*(1), 83–93. doi:10.1006/jmbi.1998.2443

Johns, T. G., Adams, T. E., Cochran, J. R., Hall, N. E., Hoyne, P. A., Olsen, M. J., … Scott, A. M. (2004). Identification of the epitope for the epidermal growth factor receptor-specific monoclonal antibody 806 reveals that it preferentially recognizes an untethered form of the receptor. *The Journal of Biological Chemistry*, *279*(29), 30375–30384.

Johns, T. G., Mellman, I., Cartwright, G. A., Ritter, G., Old, L. J., Burgess, A. W., & Scott, A. M. (2005). The antitumor monoclonal antibody 806 recognizes a high- mannose form of the EGF receptor that reaches the cell surface when cells over-express the receptor, *18*, 1–18.

Johnson, L., & Phillips, D. C. (1965). Structure of some crystalline lysozyme-inhibitor complexes determined by X-ray analysis at 6 Angstrom resolution. *Nature*, *206*(4986), 761–3.

Jorgensen, W. L., Maxwell, D. S., & Tirado-Rives, J. (1996). Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *Journal of the American Chemical Society*, *118*(45), 11225–11236. doi:10.1021/ja9621760

Jorgensen, W. L., & Tirado-rives, J. (1996). Monte Carlo vs Molecular Dynamics for Conformational Sampling, *3654*(96), 14508–14513.

Jungbluth, A. a, Stockert, E., Huang, H. J. S., Collins, V. P., Coplan, K., Iversen, K., … Cavanee, W. K. (2003). A monoclonal antibody recognizing human cancers with amplification/overexpression of the human epidermal growth factor receptor. *Proceedings of the National Academy of Sciences*, *100*(2), 639–44. doi:10.1073/pnas.232686499

Kabsch, W. (1978). A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*. doi:10.1107/S0567739478001680

Karplus, M. (1963). Vicinal Proton Coupling in Nuclear Magnetic Resonance. *Journal of the American Chemical Society*, *85*(18), 2870–2871. doi:10.1021/ja00901a059

Karplus, M., & Kuriyan, J. (2005). Molecular dynamics and protein function. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(19), 6679–85. doi:10.1073/pnas.0408930102

Karplus, M., & McCammon, J. A. (2002). Molecular dynamics simulations of biomolecules. *Nature structural biology*, *9*(9), 646–52. doi:10.1038/nsb0902-646

Katz, W. S., Lesa, G. M., Yannoukakos, D., Clandinin, T. R., Schlessinger, J., & Sternberg, P. W. (1996). A Point Mutation in the Extracellular Domain Activates LET-23 , the Caenorhabditis elegans Epidermal Growth Factor Receptor Homolog. *Molecular and Cellular Biology*, *16*(2), 529–537.

Kawabata, T. (2008). Multiple subunit fitting into a low-resolution density map of a macromolecular complex using a gaussian mixture model. *Biophysical journal*, *95*(10), 4643–58. doi:10.1529/biophysj.108.137125

Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G., Wyckoff, H., & Phillips, D. C. (1958). A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, *181*(4610), 662–666.

Kitao, a, Hayward, S., & Go, N. (1994). Comparison of normal mode analyses on a small globular protein in dihedral angle space and Cartesian coordinate space. *Biophysical chemistry*, *52*(2), 107–14.

Klepeis, J. L., Lindorff-Larsen, K., Dror, R. O., & Shaw, D. E. (2009). Long-timescale molecular dynamics simulations of protein structure and function. *Current Opinion in Structural Biology*, *19*(2), 120–127.

Kondrashov, D. a, Cui, Q., & Phillips, G. N. (2006). Optimization and evaluation of a coarse-grained model of protein motion using x-ray crystal data. *Biophysical journal*, *91*(8), 2760–7. doi:10.1529/biophysj.106.085894

Kondrashov, D. a, Van Wynsberghe, A. W., Bannen, R. M., Cui, Q., & Phillips, G. N. (2007). Protein structural variation in computational models and crystallographic data. *Structure (London, England : 1993)*, *15*(2), 169–77. doi:10.1016/j.str.2006.12.006

Koshland, D. E. (1958). Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proceedings of the National Academy of Sciences of the United States of America*, *44*(2), 98–104.

Kovacs, J. a, Chacón, P., & Abagyan, R. (2004). Predictions of protein flexibility: first-order measures. *Proteins*, *56*(4), 661–8. doi:10.1002/prot.20151

Kozer, N., Rothacker, J., Burgess, A. W., Nice, E. C., & Clayton, A. H. a. (2011). Conformational dynamics in a truncated epidermal growth factor receptor ectodomain. *Biochemistry*, *50*(23), 5130–9. doi:10.1021/bi200095w

Krall, J. a, Beyer, E. M., & MacBeath, G. (2011). High- and low-affinity epidermal growth factor receptor-ligand interactions activate distinct signaling pathways. *PloS one*, *6*(1), e15945. doi:10.1371/journal.pone.0015945

Krebs, W. G., Alexandrov, V., Wilson, C. a, Echols, N., Yu, H., & Gerstein, M. (2002). Normal mode analysis of macromolecular motions in a database framework: developing mode concentration as a useful classifying statistic. *Proteins*, *48*(4), 682–95. doi:10.1002/prot.10168

Kundu, S., Melton, J. S., Sorensen, D. C., & Phillips, G. N. (2002). Dynamics of proteins in crystals: comparison of experiment with simple models. *Biophysical journal*, *83*(2), 723–32. doi:10.1016/S0006-3495(02)75203-X

Langevin, P. (1908). Sur la théorie du mouvement brownien. *Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences*, *146*, 508–533.

Lazebnik, Y. (2002). Can a biologist fix a radio?--Or, what I learned while studying apoptosis. *Cancer cell*. doi:10.1016/S1535-6108(02)00133-2

Lee, J. C., Vivanco, I., Beroukhim, R., Huang, J. H. Y., Feng, W. L., DeBiasi, R. M., … Mellinghoff, I. K. (2006). Epidermal growth factor receptor activation in glioblastoma through novel missense mutations in the extracellular domain. *PLoS medicine*, *3*(12), e485. doi:10.1371/journal.pmed.0030485

Leioatts, N., Romo, T. D., & Grossfield, A. (2012). Elastic Network Models Are Robust to Variations in Formalism. *Journal of Chemical Theory and Computation*, *37*(7), 120613142530002. doi:10.1021/ct3000316

Levinthal, C. (1969). How to fold graciously. In J. T. P. DeBrunner & E. Munck (Eds.), *Mossbauer Spectroscopy in Biological Systems Proceedings of a meeting held at Allerton House Monticello Illinois* (Vol. 67, pp. 22–24). University of Illinois Press.

Levitt, M. (2001). The birth of computational structural biology. *Nature structural biology*, *8*(5), 392–3. doi:10.1038/87545

Li, G., & Cui, Q. (2002). A coarse-grained normal mode approach for macromolecules: an efficient implementation and application to Ca(2+)-ATPase. *Biophysical Journal*, *83*(5), 2457–2474.

Lindahl, E., & Delarue, M. (2005). Refinement of docked protein-ligand and protein-DNA structures using low frequency normal mode amplitude optimization. *Nucleic acids research*, *33*(14), 4496–506. doi:10.1093/nar/gki730

Lindemann, F. A. (1910). The calculation of molecular vibration frequencies. *Zeitschrift fur Physik*, *11*, 609–612.

Linderstrom-Lang, K. U., & Schellmann, J. A. (1959). *Protein Structure and enzyme activity. In: The Enzymes. Vol.1.* (P. D. Boyer, Ed.) (Second., pp. 443–510). New York: Academic Press.

Lindorff-Larsen, K., Piana, S., Palmo, K., Maragakis, P., Klepeis, J. L., Dror, R. O., & Shaw, D. E. (2010). Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins*, *78*(8), 1950–1958.

Liu, P., Cleveland, T. E., Bouyain, S., Byrne, P. O., Longo, P. A., & Leahy, D. J. (2012). A single ligand is sufficient to activate EGFR dimers. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(27), 10861–6. doi:10.1073/pnas.1201114109

Lopéz-Blanco, J. R., & Chacón, P. (n.d.). iMODFIT: Efficient and robust flexible fitting based on vibrational analysis in internal coordinates. *Journal of Structural Biology*, (0). doi:http://dx.doi.org/10.1016/j.jsb.2013.08.010

Lopéz-Blanco, J. R., Garzón, J. I., & Chacón, P. (2011). iMod: multipurpose normal mode analysis in internal coordinates. *Bioinformatics*, *27*(20), 2843–2850. doi:10.1093/bioinformatics/btr497

Lu, C., Mi, L., Walz, T., Springer, T. A., & Avenue, L. (2012). Mechanisms for kinase-mediated dimerization of EGFR. *Journal of Biological Chemistry*, *in press*. doi:10.1074/jbc.M112.414391

Lu, C., Mi, L.-Z., Grey, M. J., Zhu, J., Graef, E., Yokoyama, S., & Springer, T. A. (2010). Structural Evidence for Loose Linkage between Ligand Binding and Kinase Activation in the Epidermal Growth Factor Receptor. *Molecular and Cellular Biology*, *30*(22), 5432–5443.

Lu, M., Poon, B., & Ma, J. (2006). A New Method for Coarse-Grained Elastic Normal-Mode Analysis. *Journal of chemical theory and computation*, *2*(3), 464–471. doi:10.1021/ct050307u

Ma, B., Tsai, C.-J., Haliloğlu, T., & Nussinov, R. (2011). Dynamic allostery: linkers are not merely flexible. *Structure (London, England : 1993)*, *19*(7), 907–17. doi:10.1016/j.str.2011.06.002

Ma, J. (2005). Usefulness and limitations of normal mode analysis in modeling dynamics of biomolecular complexes. *Structure (London, England : 1993)*, *13*(3), 373–80. doi:10.1016/j.str.2005.02.002

MacKerell, A. D., Bashford, D., Dunbrack, R. L., Evanseck, J. D., Field, M. J., Fischer, S., … Karplus, M. (1998). All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *The Journal of Physical Chemistry B*, *102*(18), 3586–3616. doi:10.1021/jp973084f

Magnusson, U., Salopek-Sondi, B., Luck, L. A., & Mowbray, S. L. (2004). X-ray structures of the leucine-binding protein illustrate conformational changes and the basis of ligand specificity. *The Journal of biological chemistry*, *279*(10), 8747–8752. doi:10.1074/jbc.M311890200

Marrink, S. J., & Tieleman, D. P. (2013). Perspective on the Martini model. *Chemical Society reviews*. doi:10.1039/c3cs60093a

McCammon, J A, Gelin, B. R., & Karplus, M. (1977). Dynamics of folded proteins. *Nature*, *267*(5612), 585–590.

McCammon, J.A., & Harvey, S. C. (1987). *Dynamics of Proteins and Nucleic Acids*. Cambridge, UK: Cambridge University Press.

Mendez, R., & Bastolla, U. (2010). Torsional network model: normal modes in torsion angle space better correlate with conformation changes in proteins. *Physical review letters*, *104*(22), 228103. doi:10.1103/PhysRevLett.104.228103

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., & Teller, A. H. (1953). Equation of State Calculations by Fast Computing Machines. *Journal of Chemical Physics*, *21*(6), 1087–1092. doi:10.1063/1.1699114

Meyer, T., D'Abramo, M., Hospital, A., Rueda, M., Ferrer-Costa, C., Pérez, A., … Orozco, M. (2010). MoDEL (Molecular Dynamics Extended Library): a database of atomistic molecular dynamics trajectories. *Structure London England 1993*, *18*(11), 1399–1409.

Mi, L.-Z., Grey, M. J., Nishida, N., Walz, T., Lu, C., & Springer, T. A. (2008). Functional and structural stability of the epidermal growth factor receptor in detergent micelles and phospholipid nanodiscs. *Biochemistry*, *47*(39), 10314–10323.

Monod, J., Wyman, J., & Changeux, J. P. (1965). On the nature of allosteric transitions: a plausible model. *Journal of Molecular Biology*, *12*(December), 88–118.

Monticelli, L., & Tieleman, D. P. (2013). Force fields for classical molecular dynamics. *Methods in molecular biology Clifton NJ*, *924*, 197–213. doi:10.1007/978-1-62703-017-5_8

Naganathan, A. N., & Orozco, M. (2011). The native ensemble and folding of a protein molten-globule: functional consequence of downhill folding. *Journal of the American Chemical Society*, *133*(31), 12154–12161.

Navizet, I., Cailliez, F., & Lavery, R. (2004). Probing protein mechanics: residue-level properties and their use in defining domains. *Biophysical journal*, *87*(3), 1426–1435.

Navizet, I., Lavery, R., & Jernigan, R. L. (2004). Myosin flexibility: structural domains and collective vibrations. *Proteins*, *54*(3), 384–93. doi:10.1002/prot.10476

Newman, M. E. (2001). Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physical Review E - Statistical, Nonlinear and Soft Matter Physics*, *64*(1 Pt 2), 016132.

Newman, M. E. J. (2002). Assortative mixing in networks. *Physical Review Letters*, *89*(20), 5.

Newman, M. E. J. (2003). The structure and function of complex networks. (S. N. Dorogovstev & J. F. F. Mendes, Eds.)*SIAM Review*, *45*(2), 58.

Newman, M. E. J. (2004). Analysis of weighted networks. *Physical Review E*, *70*(5), 9.

Noy, A., Luque, F. J., & Orozco, M. (2008). Theoretical Analysis of Antisense Duplexes: Determinants of the RNase H Susceptibility. *Journal of the American Chemical Society*, *130*(11), 3486–3496. doi:10.1021/ja076734u

Ogiso, H., Ishitani, R., Nureki, O., Fukai, S., Yamanaka, M., Kim, J.-H., … Yokoyama, S. (2002). Crystal structure of the complex of human epidermal growth factor and receptor extracellular domains. *Cell*, *110*(6), 775–87.

Orellana, L., Carrillo, O., & Orozco, M. (2014). Large-scale Conformational Transition Pathways by Elastic Network-Langevin Dynamics. *(in preparation)*.

Orellana, L., Lopéz-Blanco, J. R., Chacón, P., & Orozco, M. (2014). Exploring the link between residue network properties and protein dynamical behavior by fast dihedral NMAe.

Orellana, L., Rueda, M., Ferrer-Costa, C., Lopez-Blanco, J. R., Chacón, P., & Orozco, M. (2010). Approaching Elastic Network Models to Molecular Dynamics Flexibility. *Journal of Chemical Theory and Computation*, *6*(9), 2910–2923. doi:10.1021/ct100208e

Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., & Thornton, J. M. (1997). CATH--a hierarchic classification of protein domain structures. *Structure*, *5*(8), 1093–1108. doi:10.1016/S0969-2126(97)00260-8

Orozco, M., Orellana, L., Hospital, A., Naganathan, A. N., Emperador, A., Carrillo, O., & Gelpí, J. L. (2011). Coarse-grained Representation of Protein Flexibility. Foundations, Successes, and Shortcomings. *Advances in protein chemistry and structural biology*, *85*, 183–215.

Özcan, F., Klein, P., Lemmon, M. A., Lax, I., & Schlessinger, J. (2006). On the nature of low- and high-affinity EGF receptors on living cells. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(15), 5735–5740.

Pasi, M., Lavery, R., & Ceres, N. (2012). PaLaCe: A Coarse-Grain Protein Model for Studying Mechanical Properties. *Journal of Chemical Theory and Computation*, *9*(1), 785–793. doi:10.1021/ct3007925

Pauling, L. (1935). The Oxygen Equilibrium of Hemoglobin and Its Structural Interpretation. *Proceedings of the National Academy of Sciences*, *21*(4), 186–191.

Pérez, A., Blas, J. R., Rueda, M., López-Bes, J. M., De La Cruz, X., & Orozco, M. (2005). Exploring the Essential Dynamics of B-DNA. *Journal of Chemical Theory and Computation*, *1*(5), 790–800. doi:10.1021/ct050051s

Periole, X., & Marrink, S.-J. (2013). The Martini coarse-grained force field. *Methods in molecular biology Clifton NJ*, *924*, 533–65. doi:10.1007/978-1-62703-017-5_20

Petrone, P., & Pande, V. S. (2006). Can conformational change be described by only a few normal modes? *Biophysical journal*, *90*(5), 1583–93. doi:10.1529/biophysj.105.070045

Piazza, F., & Sanejouand, Y.-H. (2008). Discrete breathers in protein structures. *Physical biology*, *5*(2), 026001.

Pike, L. J. (2012). Negative co-operativity in the EGF receptor. *Biochemical Society Transactions*, *40*(1), 15–9. doi:10.1042/BST20110610

Ponder, J. W., & Case, D. A. (2003). Force fields for protein simulations. *Advances in protein chemistry*, *66*, 27–85. doi:10.1016/S0065-3233(03)66002-X

Proctor, E. A., Ding, F., & Dokholyan, N. V. (2011). Discrete molecular dynamics. *Wiley Interdisciplinary Reviews Computational Molecular Science*, *1*(1), 80–92. doi:10.1002/wcms.4

Purcell, E. M., Torrey, H. C., & Pound, R. V. (1946). Resonance absorption by nuclear magnetic moments in a solid. *Physical Review*, *69*(1-2), 37–38.

Putnam, C. D., Hammel, M., Hura, G. L., & Tainer, J. A. (2007). X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution. *Quarterly Reviews of Biophysics*, *40*(03), 191–285.

Rahman, A. (1964). Correlations in the Motion of Atoms in Liquid Argon. *Physical Review*, *136*(2A), A405–A411. doi:10.1103/PhysRev.136.A405

Reif, B., Hennig, M., & Griesinger, C. (1997). Direct measurement of angles between bond vectors in high-resolution NMR. *Science (New York, N.Y.)*, *276*(5316), 1230–1233. doi:10.1126/science.276.5316.1230

Riccardi, D., Cui, Q., & Phillips, G. N. (2010). Evaluating elastic network models of crystalline biological molecules with temperature factors, correlated motions, and diffuse x-ray scattering. *Biophysical Journal*, *99*(8), 2616–2625.

Romo, T. D., & Grossfield, A. (2011). Validating and improving elastic network models with molecular dynamics simulations. *Proteins*, *79*(1), 23–34. doi:10.1002/prot.22855

Rotkiewicz, P., & Skolnick, J. (2008). Fast Procedure for Reconstruction of Full-Atom Protein Models from Reduced Representations, 1–6. doi:10.1002/jcc

Rouse, P. E. (1953). A Theory of the Linear Viscoelastic Properties of Dilute Solutions of Coiling Polymers. *The Journal of Chemical Physics*, *21*(7), 1272. doi:10.1063/1.1699180

Rueda, M., Bottegoni, G., & Abagyan, R. (2009). Consistent improvement of cross-docking results using binding site ensembles generated with elastic network normal modes. *Journal of chemical information and modeling*, *49*(3), 716–25. doi:10.1021/ci8003732

Rueda, M., Chacón, P., & Orozco, M. (2007). Thorough validation of protein normal mode analysis: a comparative study with essential dynamics. *Structure*, *15*(5), 565–75. doi:10.1016/j.str.2007.03.013

Rueda, M., Ferrer-Costa, C., Meyer, T., Pérez, A., Camps, J., Hospital, A., … Orozco, M. (2007a). A consensus view of protein dynamics. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(3), 796–801. doi:10.1073/pnas.0605534104

Rueda, M., Ferrer-Costa, C., Meyer, T., Pérez, A., Camps, J., Hospital, A., … Orozco, M. (2007b). A consensus view of protein dynamics. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(3), 796–801. doi:10.1073/pnas.0605534104

Rueda, M., Ferrer-Costa, C., Meyer, T., Pérez, A., Camps, J., Hospital, A., … Orozco, M. (2007c). A consensus view of protein dynamics. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(3), 796–801.

Rzepiela, A. J., Louhivuori, M., Peter, C., & Marrink, S. J. (2011). Hybrid simulations: combining atomistic and coarse-grained force fields using virtual sites. *Physical chemistry chemical physics : PCCP*, *13*(22), 10437–48.

Sacquin-Mora, S., & Lavery, R. (2006). Investigating the local flexibility of functional residues in hemoproteins. *Biophysical journal*, *90*(8), 2706–2717.

Saunders, M. G., & Voth, G. A. (2013). Coarse-Graining Methods for Computational Biology. *Annual review of biophysics*. doi:10.1146/annurev-biophys-083012-130348

Schiefner, A., Holtmann, G., Diederichs, K., Welte, W., & Bremer, E. (2004). Structural basis for the binding of compatible solutes by ProX from the hyperthermophilic archaeon Archaeoglobus fulgidus. *The Journal of biological chemistry*, *279*(46), 48270–48281. doi:10.1074/jbc.M403540200

Schlick, T. (2001). Time-Trimming Tricks for Dynamic Simulations: Splitting Force Updates to Reduce Computational Work. *Structure*, *9*(4), R45–R53. doi:http://dx.doi.org/10.1016/S0969-2126(01)00593-7

Schneider, M. R., & Wolf, E. (2009). The epidermal growth factor receptor ligands at a glance. *Journal of cellular physiology*, *218*(3), 460–6. doi:10.1002/jcp.21635

Schrödinger, E. (1992). *What is life?: The physical aspect of the living cell ; with Mind and matter ; & Autobiographical sketches*. *Mind and Matter* (p. 204). Cambridge University Press.

Sen, T. Z., & Jernigan, R. L. (2006). Optimizing the Parameters of the Gaussian Network Model for ATP-Binding Proteins. In *Normal Mode Analysis. Theory and Applications to Biological and Chemical Systems» Chapman&Hall/CRC. Mathematical and Computational Biology Series. Ed. Qiang Cui & Ivet Bahar* (pp. 171–186).

Senn, H. M., & Thiel, W. (2009). QM/MM methods for biomolecular systems. *Angewandte Chemie (International ed. in English)*, *48*(7), 1198–1229.

Sfriso, P., Emperador, A., Orellana, L., Hospital, A., Gelpí, J. L., & Orozco, M. (2012). Finding conformational Transition Pathways from Discrete Molecular Dynamics Simulations. *Journal of Chemical Theory and Computation*.

Shan, Y., Arkhipov, A., Kim, E. T., Pan, A. C., & Shaw, D. E. (2013). Transitions to catalytically inactive conformations in EGFR kinase. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(18), 7270–5. doi:10.1073/pnas.1220843110

Shan, Y., Eastwood, M. P., Zhang, X., Kim, E. T., Arkhipov, A., Dror, R. O., … Shaw, D. E. (2012). Oncogenic Mutations Counteract Intrinsic Disorder in the EGFR Kinase and Promote Receptor Dimerization. *Cell*. doi:10.1016/j.cell.2012.02.063

Shaw, D. E., Bowers, K. J., Chow, E., Eastwood, M. P., Ierardi, D. J., Klepeis, J. L., … Batson, B. (2009). Millisecond-scale molecular dynamics simulations on Anton. *Proceedings of the Conference on High Performance Computing Networking Storage and Analysis SC 09*, (c), 1. doi:10.1145/1654059.1654099

Shaw, D. E., Maragakis, P., Lindorff-Larsen, K., Piana, S., Dror, R. O., Eastwood, M. P., … Wriggers, W. (2010). Atomic-Level Characterization of the Structural Dynamics of Proteins. *Science* , *330* (6002), 341–346. doi:10.1126/science.1187409

Shirvanyants, D., Ding, F., Tsao, D., Ramachandran, S., & Dokholyan, N. V. (2012). Discrete Molecular Dynamics: An Efficient And Versatile Simulation Method For Fine Protein Characterization. *The Journal of Physical Chemistry B*, 1–15. doi:10.1021/jp2114576

Stillinger, F. H., & Rahman, A. (1974). Improved simulation of liquid water by molecular dynamics. *Journal of Chemical Physics*, *60*, 1545.

Strogatz, S. H. (2001). Exploring complex networks. *Nature*, *410*(6825), 268–276.

Suhre, K., Navaza, J., & Sanejouand, Y. H. (2006). NORMA: a tool for flexible fitting of high-resolution protein structures into low-resolution electron-microscopy-derived density maps. *Acta Crystallographica Section D Biological Crystallography*, *62*(Pt 9), 1098–1100.

Suhre, K., & Sanejouand, Y.-H. (2004). ElNemo: a normal mode web server for protein movement analysis and the generation of templates for molecular replacement. *Nucleic acids research*, *32*(Web Server issue), W610–4. doi:10.1093/nar/gkh368

Sun, S., Chandler, D., Dinner, A. R., & Oster, G. (2003). Elastic energy storage in beta-sheets with application to F1-ATPase. *European biophysics journal : EBJ*, *32*(8), 676–683. doi:10.1007/s00249-003-0335-6

Sutto, L., & Gervasio, F. L. (2013). Effects of oncogenic mutations on the conformational free-energy landscape of EGFR kinase. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(26), 10616–21. doi:10.1073/pnas.1221953110

Tama, F, Gadea, F. X., Marques, O., & Sanejouand, Y. H. (2000). Building-block approach for determining low-frequency normal modes of macromolecules. *Proteins*, *41*(1), 1–7.

Tama, F, & Sanejouand, Y. H. (2001). Conformational change of proteins arising from normal mode calculations. *Protein engineering*, *14*(1), 1–6.

Tama, Florence, Miyashita, O., & Brooks, C. L. (2004). Normal mode based flexible fitting of high-resolution structure into low-resolution experimental data from cryo-EM. *Journal of Structural Biology*, *147*(3), 315–326.

Teramura, Y., Ichinose, J., Takagi, H., Nishida, K., Yanagida, T., & Sako, Y. (2006). Single-molecule analysis of epidermal growth factor binding on the surface of living cells. *the The European Molecular Biology Organization Journal*, *25*(18), 4215–4222.

Tirion, M. (1996). Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. *Physical review letters*, *77*(9), 1905–1908.

Tolonen, E., Bueno, B., Kulshreshta, S., Cieplak, P., Argáez, M., Velázquez, L., & Stec, B. (2011). Allosteric transition and binding of small molecule effectors causes curvature change in central β-sheets of selected enzymes. *Journal of molecular modeling*, *17*(4), 899–911. doi:10.1007/s00894-010-0784-7

Tozzini, V. (2005). Coarse-grained models for proteins. *Current opinion in structural biology*, *15*, 144–150. doi:10.1016/j.sbi.2005.02.005

Ultsch, M. H., Wiesmann, C., Simmons, L. C., Henrich, J., Yang, M., Reilly, D., … de Vos, A. M. (1999). Crystal structures of the neurotrophin-binding domain of TrkA, TrkB and TrkC. *Journal of Molecular Biology*, *290*(1), 149–159. doi:http://dx.doi.org/10.1006/jmbi.1999.2816

Van Wynsberghe, A. W., & Cui, Q. (2006). Interpreting correlated motions using normal mode analysis. *Structure*, *14*(11), 1647–53. doi:10.1016/j.str.2006.09.003

Velázquez-Muriel, J. A., Rueda, M., Cuesta, I., Pascual-Montano, A., Orozco, M., & Carazo, J.-M. (2009). Comparison of molecular dynamics and superfamily spaces of protein domain deformation. *BMC Structural Biology*, *9*, 6.

Vendruscolo, M., Dokholyan, N., Paci, E., & Karplus, M. (2002). Small-world view of the amino acids that play a key role in protein folding. *Physical Review E*, *65*(6), 1–4. doi:10.1103/PhysRevE.65.061910

Verlet, L. (1967). Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Physical Review*. doi:10.1103/PhysRev.159.98

Vishveshwara, S., Brinda, K. V, & Kannan, N. (2002). PROTEIN STRUCTURE : INSIGHTS FROM GRAPH THEORY, *1*(1), 1–25.

Vivanco, I., Robins, H. I., Rohle, D., Campos, C., Grommes, C., Nghiemphu, P. L., … Mellinghoff, I. K. (2012). Differential Sensitivity of Glioma- versus Lung Cancer-Specific EGFR Mutations to EGFR Kinase Inhibitors. *Cancer Discovery*, *2*(5), 458–471. doi:10.1158/2159-8290.CD-11-0284

Vögeli, B., Yao, L., & Bax, A. (2008). Protein backbone motions viewed by intraresidue and sequential HN-Halpha residual dipolar couplings. *Journal of biomolecular NMR*, *41*(1), 17–28.

Walker, F., Orchard, S. G., Jorissen, R. N., Hall, N. E., Zhang, H.-H., Hoyne, P. A., … Burgess, A. W. (2004). CR1/CR2 interactions modulate the functions of the cell surface epidermal growth factor receptor. *The Journal of Biological Chemistry*, *279*(21), 22387–22398.

Weiss, D. R., & Levitt, M. (2009). Can Morphing Methods Predict Intermediate Structures? *Journal of Molecular Biology*, *385*(2), 665–674. doi:http://dx.doi.org/10.1016/j.jmb.2008.10.064

Wiederstein, M., & Sippl, M. J. (2007). ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic acids research*, *35*(Web Server issue), W407–W410.

Wieduwilt, M. J., & Moasser, M. M. (2008). The epidermal growth factor receptor family: biology driving targeted therapeutics. *Cellular and molecular life sciences CMLS*, *65*(10), 1566–1584.

Williamson, M. P., Havel, T. F., & Wüthrich, K. (1985). Solution conformation of proteinase inhibitor IIA from bull seminal plasma by 1H nuclear magnetic resonance and distance geometry. *Journal of molecular biology*, *182*(2), 295–315. doi:10.1016/0022-2836(85)90347-X

Woolf, T. B. (1998). Path corrected functionals of stochastic trajectories: towards relative free energy and reaction coordinate calculations. *Chemical Physics Letters*, *289*(5–6), 433–441. doi:http://dx.doi.org/10.1016/S0009-2614(98)00427-8

Wuethrich, K. (1989). The development of nuclear magnetic resonance spectroscopy as a technique for protein structure determination. *Accounts of Chemical Research*, *22*(1), 36–44. doi:10.1021/ar00157a006

Wüthrich, K. (2001). The way to NMR structures of proteins. *Journal of Molecular Biology*, *8*(11), 923–925. doi:10.1038/nsb1101-923

Xia, F., Tong, D., & Lu, L. (2013). Robust Heterogeneous Anisotropic Elastic Network Model Precisely Reproduces the Experimental B-factors of Biomolecules. *Journal of Chemical Theory and Computation*, *9*(8), 3704–3714. doi:10.1021/ct4002575

Xu, Y., Moseley, J. B., Sagot, I., Poy, F., Pellman, D., Goode, B. L., & Eck, M. J. (2004, March 5). Crystal Structures of a Formin Homology-2 Domain Reveal a Tethered Dimer Architecture. *Cell*. Cell Press.

Yang, L., Song, G., Carriquiry, A., & Jernigan, R. L. (2008). Close correspondence between the motions from principal component analysis of multiple HIV-1 protease structures and elastic network modes. *Structure (London, England : 1993)*, *16*(2), 321–30. doi:10.1016/j.str.2007.12.011

Yang, L., Song, G., & Jernigan, R. L. (2009). Protein elastic network models and the ranges of cooperativity. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(30), 12347–12352.

Yang, L.-W., Eyal, E., Bahar, I., & Kitao, A. (2009a). Principal component analysis of native ensembles of biomolecular structures (PCA_NEST): insights into functional dynamics. *Bioinformatics*, *25*(5), 606–614.

Yang, L.-W., Eyal, E., Bahar, I., & Kitao, A. (2009b). Principal component analysis of native ensembles of biomolecular structures (PCA_NEST): insights into functional dynamics. *Bioinformatics*, *25*(5), 606–14. doi:10.1093/bioinformatics/btp023

Yang, L.-W., Eyal, E., Chennubhotla, C., Jee, J., Gronenborn, A. M., & Bahar, I. (2007). Insights into equilibrium dynamics of proteins from comparison of NMR and X-ray data with computational predictions. *Structure (London, England : 1993)*, *15*(6), 741–9. doi:10.1016/j.str.2007.04.014

Yang, Z., Májek, P., & Bahar, I. (2009). Allosteric transitions of supramolecular systems explored by network models: application to chaperonin GroEL. *PLoS computational biology*, *5*(4), e1000360. doi:10.1371/journal.pcbi.1000360

Yesylevskyy, S. O., Schäfer, L. V, Sengupta, D., & Marrink, S. J. (2010). Polarizable Water Model for the Coarse-Grained MARTINI Force Field. (Michael Levitt, Ed.)*PLoS Computational Biology*, *6*(6), 17.

Zacharias, M. (2003). Protein-protein docking with a reduced protein model accounting for side-chain flexibility. *Protein science : a publication of the Protein Society*, *12*(6), 1271–1282.

Zandi, R., Larsen, A. B., Andersen, P., Stockhausen, M.-T., & Poulsen, H. S. (2007). Mechanisms for oncogenic activation of the epidermal growth factor receptor. *Cellular signalling*, *19*(10), 2013–23. doi:10.1016/j.cellsig.2007.06.023

Zheng, W., & Doniach, S. (2003). A comparative study of motor-protein motions by using a simple elastic-network model. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(23), 13253–13258.

Zhou, Y., Vitkup, D., & Karplus, M. (1999). Native proteins are surface-molten solids: application of the Lindemann criterion for the solid versus liquid state. *Journal of molecular biology*, *285*(4), 1371–5. doi:10.1006/jmbi.1998.2374

Zuckerman, D. M., & Woolf, T. B. (1999). Dynamic reaction paths and rates through importance-sampled stochastic dynamics. *The Journal of Chemical Physics*, *111*(21).

Zuckerman, D. M., & Woolf, T. B. (2000). Efficient dynamic importance sampling of rare events in one dimension. *Physical Review E*, *63*(1), 16702.

Acknowledgments

First of all, I would like to thank my advisor, Dr. Modesto Orozco, for giving me an opportunity, for always trusting in me and giving me freedom to pursue my own ideas, and above all, being human when needed. To all the people in the lab, starting by the old generation of desayunos@, the regulars of *"La Terrasseta"* at Fridays I had the pleasure to meet when I arrived. Agnes, for welcoming me in my first day and helping me to switch on a computer – your hairdryer advice for babies was very useful too. Agustí and Oliver, for lots of laughs together, discussions, tons of help, and also for increasing even more my love and hunger for physics (and coarse-graining). Manu, for transferring me the basics of NMA during my first steps. Annalisa, for being a friend during these years, for many lunchs, and coffees and talks. I hope the end of our thesis will be the start of a long long friendship. Adam and Jose, for being always helpful in spite of your nightmare tasks - Jose, I will always remember how you resurrected my *"barrulaptop"* and how you take care of my back, and Adam, infinite thanks for helping me to finish my dearest project and being enthusiastic on it. Montse and Rebeca, for your warmth and kindness. Tim and Guillem, thanks for being always ready to help. To Bryn, for bringing me an enriching chance to bring your experimental world and fine details I used not to pay attention to my work. To Marga, for all your administrative help and also many nice conversations. To Pedro, thanks for strikingly original insights and fruitful brainstorming. Last but not least, to all the people that I have met in the lab, for your companionship and kindness: Nacho, Michaela, Rosana, Federica, Oscar, Pablo, Andreu, Nadine, Rima, Antonella...It has been an enriching experience to spend these years in such a multicultural and diverse environment. In addition, I would like to thank Prof. Ignasi Fita and Prof. Xavier de La Cruz, for his support and advice. The small sectarian group of NMA in Madrid, Pablo Chacón and José Ramón López-Blanco deserves a special mention. Thanks for enormous patience in the development of our collaboration. Mon, in spite of the distance, you have been a friend and always helpful; thanks you for your great hospitality in my two trips to Madrid, especially in your messy pre-thesis flat in Alcorcón.

Simone and Lou, for wonderful dinners and moments together, since those old times at *carrer Napols*. A mis fantásticos tíos Antonio y Loli, por ayudarnos con el huerto y con la peque respectivamente. A tots els meus amics animals: Llamp, Joy, Luna, Cuca, Lola, Titu, que m'han acompanyat al llarg dels anys.

I could also thank the great science writer Carl Sagan, whom I had loved to meet and that was in some part, responsible of getting me into this. Reading the book *Cosmos* at age 11 was almost a mystical experience that made me fell in love with all I am doing today – Biology, Physics, SEARCH, DISCOVER.

And last but above all. A la meva mare, Remei. *Gràcies per retornar; he lluitat per tots nosaltres.* And to Johan: without you, everything, not only this thesis, but also all the good things that flourished along with it, would have never been possible. This work has been for me the culmination of a very long path of struggle, which began when I was a child, fascinated by knowledge and with many odds against it. Now, I feel I have achieved at last the top of a mountain I started to climb years ago, not only scientifically, but as human being. The climbing has been though, the paths tortuous at times, luminous others, but now I feel my lungs full with fresh air and energy, and start to enjoy the views – many other mountains to reach. To end: Thanks to the power that has driven me through the years. As someone told me once: ***"life is sometimes nothing but a muddle, but it also forms a beautiful and authentic picture when one sees it from a distance".***

This page intentionally left blank

If you can wait and not be tired by waiting.

Rudyard Kipling