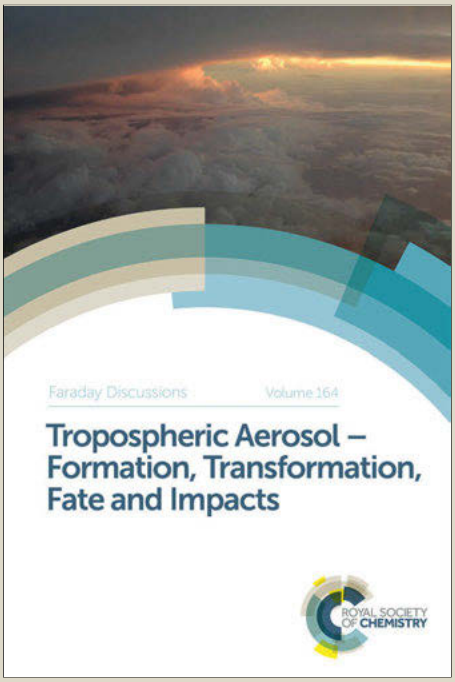
Faraday Discussions

Accepted Manuscript



Faraday Discussions Volume 164

Tropospheric Aerosol – Formation, Transformation, Fate and Impacts

ROYAL SOCIETY OF CHEMISTRY

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/faraday_d

Allosteric Pathway Identification through Network Analysis: from Molecular Dynamics Simulations to Interactive 2D and 3D Graphs

Ariane Allain[1], Isaure Chauvot de Beauchêne[1#], Florent Langenfeld[1], Yann Guarracino[1], Elodie Laine[1$], and Luba Tchertanov[1,2*]

[1] Bioinformatics, Molecular Dynamics & Modeling (BiMoDyM), Laboratoire de Biologie et Pharmacologie Appliquée (LBPA UMR8113 CNRS), École Normale Supérieure de Cachan, 61 avenue du Président Wilson, 94235 Cachan, France

[2] Laboratoire d'excellence en Recherche sur le Médicament et l'innovation Thérapeutique (LERMIT), Campus Paris Saclay, France

# Present address: Physik-Department (T38), Technische Universität München, James-Franck-Str. 1, 85748 Garching, Germany

$ Present address: Biologie Computationnelle et Quantitative, UMR 7238 CNRS-Université Pierre et Marie Curie, 15, Rue de l'Ecole de Médecine, 75006 Paris, France

* Correspondance to: L. Tchertanov, BiMoDyM, LBPA, CNRS-ENS de Cachan, 61 Avenue du Président Wilson, 94235 Cachan, France; e-mail : Luba.Tchertanov@lbpa.ens-cachan.fr

ABSTRACT

Allostery is a universal phenomenon that couples the information induced by a local perturbation (effector) in a protein to spatially distant regulated sites. Such an event can be described in terms of a large scale transmission of information (communication) through a dynamic coupling between structurally rigid (minimally frustrated) and plastic (locally frustrated) clusters of residues. To elaborate a rational description of allosteric coupling, we propose an original approach - MOdular NETwork Analysis (MONETA) - based on the analysis of inter-residue dynamical correlations to localize the propagation of both structural and dynamical effects of a perturbation throughout a protein structure. MONETA uses inter-residue cross-correlations and commute times computed from molecular dynamics simulations and a topological description of a protein to build a modular network representation composed of clusters of residues (*dynamic segments*) linked together by chains of residues (c*ommunication pathways*). MONETA provides a brand new direct and simple visualization of protein allosteric communication. A GEPHI module implemented in MONETA package allows generating 2D graphs of communication network. An interactive PyMOL plugin permits to draw the communication pathways between chosen protein fragments or residues on a 3D representation. MONETA is a powerful tool for on-the-fly display of communication networks in proteins. We applied MONETA for analysis of communication pathways (i) between the main regulatory fragments of receptors tyrosine kinases (RTKs), KIT and CSF-1R, in the native and mutated states and (ii) in proteins STAT5(STAT5a and STAT5b) in the phosphorylated and the unphosphorylated forms. The description of the physical support for allosteric coupling by MONETA allowed a comparison of the mechanisms of (a) constitutive activation induced by equivalent mutations in two RTKs and (b) allosteric regulation in the activated and non-activated STAT5 proteins. Our theoretical prediction based on results obtained with MONETA was validated for KIT by *in vitro* experiments. MONETA is a versatile analytical and visualization tool entirely devoted to the understanding of the functioning/malfunctioning of allosteric regulation in proteins, a crucial basis to guide the discovery of next-generation allosteric drugs.

1. INTRODUCTION

Data and computational tools

With the permanently increasing computational means and the rapid development of efficient algorithms, the biomolecular modeling field has entered a new era. The exponential growth of data produced by molecular dynamics (MD) simulations calls for the development of tools able to handle and analyze these big data by reducing their dimensionality to extract relevant and pertinent information. Such tools should enable to visualize this information to unveil its biological implications and help generating new hypotheses. This issue is one of multi-scale nature. On a more systemic level, cellular spatial organization, signaling and communication also occur at different scales – from distances of some angstroms in binary protein complexes and assemblies to global micrometer distances across and between cells. Recently a new view of cellular spatial organization was proposed [1] in which cell signaling is described in terms of dynamic allosteric interactions within and among distinct, spatially organized transient clusters. The clusters vary over time and space and their lengths range from nanometers to micrometers. A key challenge is to understand the interplay across these multiple scales, link it to the physicochemical basis of the conformational behavior of single molecules and ultimately relate it to cellular function. These conceptions have emphasized the necessity for the development of computing tools for a rational understanding of allosteric regulation and its manifestations at different levels.

The last decades have faced a similar issue with the explosion of the amount of genomics and protein-protein interactions data related to cell metabolism and regulation. As a response, many performing tools were developed to integrate the data in an accurate, complete and ergonomic way. Apart from genome-wide association studies (GWAS) were developed graphical representations of gene regulatory networks (*TranscriptomeBrowser* 3.0 [2]), of networks of functionally related genes (*PainNetworks* [3]), of protein-protein interaction networks in bacteria [4], and even multi-scale and multi-object graphs of organ-disease-drug causal relationship networks [5].

Increasing interest in the allosteric phenomenon stimulated the development of tools capable to analyze and visualize protein communication at the atomic level. Some of them, as RINalyzer [6], AlloPathFinder [7] and others [8-10] build an interaction network from the native contacts observed in the crystallographic structure of a protein. SPACER [11] was developed using theoretical concepts based on the thermodynamic view of allostery. xPyder [12] and similar studies [13-15] exploit molecular-dynamics derived dynamical correlations while AD-ENM [16] and PARS [17] analyze structural communication routes based on highly simplified physical models (elastic network models or normal mode analysis). Piazza and colleagues used off-lattice and non-linear network protein models to localize modes of nonlinear origin (*discrete breathers*) that accumulate energy on few sites in response to perturbations even occurring at distant locations [18-20]. Methodological efforts have also been directed toward the identification of evolutionarily conserved networks of residues thought to mediate allosteric communication [7, 21].

The development of tools for an accurate and complete description of protein allosteric communication requires a deep discerning of the allosteric phenomenon and associated regulation mechanisms.

Allostery and associated regulation mechanisms

Allosteric regulation is an ubiquitous mechanism of intra- and intermolecular regulation of biological processes. All physiological signaling pathways implying cascades of protein-protein binding events are initiated by an intramolecular allosteric transition occurring within a given receptor in response to a specific perturbation (*e.g.*, ligand binding). As a typical paradigm, the receptor tyrosine kinases (RTKs) act as sensors for extracellular ligands, the binding of which triggers receptor dimerization and activation of the kinase function, leading to the recruitment, phosphorylation and activation of multiple downstream signaling proteins, which ultimately govern the physiology of cells [22]. RTKs allosteric activity is tightly and finely regulated under normal physiological conditions. Deregulation of RTKs activity prompts perturbation in signaling pathways and causes a wide range of diseases, particularly cancers [23].

At the molecular level, allostery is manifested as the regulation of a protein function by the binding of an effector (ligand/substrate/protein) at a site other than the active site of the protein. This mechanical perturbation triggers a specific conformational rearrangement at another spatially distant protein site. Discovered initially by the pioneering studies of the oxygen binding to hemoglobin by Monod [24], allostery is now commonly recognized as a universal phenomenon in proteins, that is intimately related to their structure, dynamics and functions. The allosteric effects have different nature and control nearly all biological processes, *e. g.,* cellular network for signal transduction, metabolism activation, motor work and transcription control. Allostery regulates activation/deactivation in enzymes, opening/closing of the ion channels, oligomerization/monomerization (dimerization of RTKs), association/dissociation of macromolecular subunits (G proteins), lost/gain of affinity for a ligand (O_2 binding to hemoglobin) or a membrane, exposure/burial of the recognition site to other protein or nucleic acid.

It is believed that proteins have the remarkable property of transmitting intramolecular signals through specific fold-encoded pathways that form a network of covalently bonded and non-covalently bonded interactions between the protein residues [24-25]. The first theories considered allosteric communication as a domino effect where local structural distortions propagate step by step through a protein structure viewed as globally static [26]. The vision of protein structure evolved from *an ensemble of discrete conformations* to a *continuous landscape of conformations*, which led to reconsider the allosteric phenomenon as a consequence of re-distributions of protein conformational ensembles (shift in conformational preferences) [27-28]. A population shift in conformational space of a given protein can be induced by the binding of an effector, point mutations or environmental conditions (pH, voltage, ions concentration…).

Allostery offers a highly specific way to modulate protein function. So far as allosteric events perturb the structure, and the energetic strain propagates and shifts the population, this can lead to changes in the shapes and properties of target binding sites. Consequently, combining equilibrium fluctuation concepts with genome-wide screens could considerably help drug discovery. However, allosteric signal transmission is difficult to detect experimentally and to model because it is often mediated by local structural changes propagating along multiple pathways. These concepts have emphasized the necessity for the development of computing tools able to account for both conformational global changes and local fluctuations to elaborate rational descriptions of allosteric coupling.

For such purpose, we propose here an original approach based on the analysis of inter-residue dynamical correlations to localize the propagation of both structural and dynamical effects of a perturbation throughout a protein structure. The principle of this approach consists in building a modular network representation of a protein, composed of clusters of residues representing *Independent Dynamic Segments* (*IDSs*) and chains of residues acting as

Communication Pathways (*CPs*) (Figure 1). The representation is derived from the topology of the protein and the inter-residue dynamical correlations computed from molecular dynamics simulations. The utilities for a modular network representation of protein structures are implemented in an interactive package MONETA (MOdular NETwork Analysis) designed for generation of the communication landscape in a protein and its visualization/ representation in 2D and 3D. These functionalities are very useful for exploring communications in allosterically regulated proteins.

2. IMPLEMENTATION and FUNCTIONALITY

The first release of MONETA (MONETA 1.0) used for the analysis of communications between two distant regions, A-loop and JMR, in the receptor tyrosine kinase KIT [29], was significantly improved, optimized and enriched by additional functionalities. Optimization was oriented principally to (i) increasing performance, allowing to significantly reduce computing time (by a factor of 30); (ii) reducing the number and volume of generated temporary files and (iii) introducing a 2D graphical representation of communication network. The advanced version of MONETA (release MONETA 1.1) [30] contains novel options for data analysis, *i.e.,* display of intra-protein connections network in a dynamical graph for the rapid identification of biologically relevant communication pathways between residues or domains. The advanced version of MONETA was used for the analysis of allosteric communication in different proteins divergent by their size, form and molecular architecture.

2. 1. MONETA composition and functionalities

MONETA analyses and extracts data from conformational ensembles computed by all-atom MD simulations of a protein to infer the topological connections (residues interactions) and the dynamical correlations between residues or domains (Figure 1).

MONETA was implemented in the form of a package that automatically performs all computational steps and data analysis – from PCA of the MD simulations to a viewing of communication landscape – through a python scripting interface to the software R [31], PyMOL [32] and Gephi [33]. MONETA running requires some additional software – the *ptraj* module of AmberTools [34], the module *g_mdmat* (or *g_mdmat_d*) of GROMACS [35], the modules

HBPLUS [36] and HBADD [36]. The package is currently composed of three parts: (i) the preparatory phase generates the data required for communication analysis from an MD trajectory and a 3D protein structure, (ii) the application phase performs the computation of intra-protein *communication pathways* and *independent dynamics segments*, (iii) the visualization phase allows data analysis and graphical representation of generated data (Figure 2).

The user gets interactive and modular representations of intra-protein connections at the protein and residue scales and of communication pathways at the atomic scale. The analysis *per se* is a three-step procedure:

(i) Identification of protein regions displaying the most striking features of the protein's local dynamics. Such regions, or *Independent Dynamics Segments*, constitute residue clusters which atomic fluctuations are highly concerted within each cluster although independent from the rest of the protein. These segments were previously described in the literature as conformationally sensitive to a disturbance of the global protein dynamics, capable of energy storage and transfer [18-19, 37]. The *IDSs* are identified by a statistical technique, Local Feature Analysis [38], adapted for analysis of the atomic coordinate fluctuations from MD simulations [39].

(ii) Detection of *communication pathways* linking the *IDS*s through non-bonded interactions between residues. The ability of the protein residues to communicate efficiently is evaluated by using the measure of communication propensity [40]. The communication between two residues is estimated by their commute time, expressed as the variance of their inter-residue distance over MD trajectories [15, 41]. Chains of residues interacting by pair and displaying high communication propensities between them would represent pathways of well-defined interactions through which information would be transmitted efficiently. Such chains of residues are denoted as *communication pathways*.

(iii) Visualization of *IDSs* and *CPs* in a communication profile of the protein. This feature of MONETA is based on the automatic computing of all *communication pathways* from the non-bonded interactions between residues and the variance of inter-atomic distances (*i.e.*, dynamic correlation) in MD simulations. An implemented GEPHI [33] module translates these connectivity groups and pathways into a 2D network graph (Figure 2). The 3D representations of *CPs* at the atomic level are performed by a PyMOL [32] module implemented in MONETA.

2.2. Theoretical background

Inter-Residue contacts (topology)

The topological description of a protein at the level of inter-residue contacts can be derived from distances matrices consisting of the average smallest distances between all residues pairs, computed using the *g_mdmat* program of GROMACS [35]. Two residues *i* and *j* are considered topologically close if the average smallest distance between them is lower than a given threshold d_{cut} = 3.6 Å. This value accounts for non-covalent contacts between protein residues such as hydrogen bonds, hydrophobic interactions and salt-bridges.

Local representation from PCA

The *IDSs* identification is based on a statistical technique, Local Feature Analysis, originally developed for image analysis [38]. LFA processes data from Principal Component Analysis and extracts local outputs of reduced dimensionality from the global PCA modes. The LFA formalism was adapted to study essential dynamics in proteins [39]. This adapted LFA was implemented in MONETA 2.0 for identification of the *independent dynamic segments.*

Identification IDSs: The PCA of a MD trajectory containing *m* conformations of a set of *N* atoms (*e.g.,* the Cα atoms of a protein composed of *N* residues) allows reconstructing the displacement D_{α_i} of a coordinate α_i over the trajectory according to the projections A_r of the trajectory on each rth mode (1). The dimensionality of the trajectory can be reduced by keeping only *n* first modes that describe the main part of the dynamics of the system (*n<<3N*) and reconstructing an approximated trajectory (2).

$$\mathrm{D}_{\alpha_i \mathrm{recons}} = \sum_{\mathrm{r}=1}^{3\mathrm{N}} \mathrm{A_r} \varphi_\mathrm{r}(i) \quad \text{with } \varphi_\mathrm{r} \text{ the r}^{\text{th}} \text{ PCA mode} \tag{1}$$

$$\mathrm{D'}_{\alpha_i \mathrm{recons}} = \sum_{\mathrm{r}=1}^{\mathrm{n}} \mathrm{A_r} \varphi_\mathrm{r}(i) \tag{2}$$

$$\mathrm{A_r} = \sum_{\mathrm{i}=1}^{'3\mathrm{N}} \varphi_\mathrm{r}(i) \mathrm{D}_{\alpha_i} = \sum_{\mathrm{i}=1}^{'3\mathrm{N}} \mathrm{K}^{\mathrm{PCA}}(\mathrm{r,i}) \mathrm{D}_{\alpha_i} \tag{3}$$

In (3), K^{PCA} represents the kernel of the transformation of the trajectory from Cartesian global coordinates consisting in the PCA modes.The PCA offers a reduced dimensionality, however, it is nonlocal. The principle of LFA is to describe movements of reduced dimensionality, not global as are the PCA modes but local. A new kernel K^{LFA} of the transformation is derived by analogy with the PCA (4).

$$K^{LFA}(j,i) = \sum_{r=1}^{'n} \varphi_r(j) \frac{1}{\sqrt{\lambda_r}} \varphi_r(i) \qquad (4)$$

Similarly to PCA outputs A_r (3), local outputs Oj are defined:

$$O_j = \sum_{r=1}^{n} \frac{A_r}{\sqrt{\lambda_r}} \varphi_r(j) \qquad (5)$$

As PCA outputs A_r are decorrelated by frequency (corresponding to the index r), LFA outputs Oj are decorrelated by space when n tends to $3N$. In the general case as $n<<3N$ LFA outputs Oj are minimally correlated. The residual correlations are expressed as (6) [38].

$$<O_i O_j> = \sum_{r=1}^{n} \varphi_r(ji)\varphi_r(j) = P(i,j) \qquad (6)$$

The $3N$ local features O_j could be computed from n PCA modes. To reduce the dimensionality of the displacement D' described by these LFA modes, the set of n features O_j which best approximates D' was selected. We refer to them as $O_{jm(m\ in\ M)}$ where M is the set of n *seeds* corresponding to the coordinates around which are centered the O_{im}. The *seeds* are obtained by applying the sparsification algorithm described in [39]. Briefly, at each step, among the ($3*n$ - $3*m$) coordinates of the atoms not in the current set M, the (m +1)th *seed* is chosen as the one which atomic fluctuations are the least described by the atomic fluctuations of the *seeds* already in the set M. An additional requirement is that the (m +1)th *seed* coordinate is not a coordinate of an atom (*i*) with another coordinate already in M or (*ii*) adjacent to an atom with a coordinate in M. This condition ensures that the new *seed* corresponds to an atom dynamically decorrelated from the *seeds* in M.

As each *IDS* describes a very high proportion of the protein fluctuations, several *seeds* can be identified at proximity of each other, although they belong to overlapping dynamic segments. To limit this phenomenon, a dimensionality reduction step was added to the LFA algorithm according to the following protocol: If two *seeds* correspond to two residues located within 6 positions from each other in the peptide sequence, only one of the two *seeds* was kept, chosen as the one describing the largest part of protein fluctuations. The LFA procedure thus identifies a number of *seeds* lower or equal to the number n of PCA modes included in the procedure.

Growth of IDSs: Each *IDS* S_m is iteratively constructed around a *seed* as a group of neighboring residues with correlated movements. A residue *i* is added to S_m if (*a*) it is topologically close to at least one residue in S_m and (*b*) its average correlation with the residues in S_m is greater than a threshold P_{cut} (7). The correlation between two residues *h* and *k* is evaluated as (8).

$$\frac{\sum_{j \epsilon S_m} P(i,j)}{|S_m|} \geq P_{cut} \qquad (7)$$

$$P(h,k) = \sum_{r=1}^{n} \sum_{d=1}^{3} \varphi_r(h_d)\, \varphi_r(k_d) \qquad (8)$$

where *d* is the (x,y,z) coordinate index and *n* is the number of retained PCA modes.

Communication pathways

CPs are chains of neighboring residues whose communication propensities between each other are high. We define them according to the concept of «*communication propensity*» [42]. The communication propensity of a pair of residues is inversely related to their *commute time CT(i, j)*, expressed as a function of the variance of the inter-residue distance (9):

$$CT(i,j) =< (d_{ij} \_ \overline{d_{ij}})^2 > \qquad (9)$$

where *dij* = |*ri* − *rj*| is the distance between the Cα atoms of residue *i* and residue *j.*

Intuitively, if a residue moves, it transmits this movement to another residue which moves in its turn, and these residues communicate all the more effectively that the delay separating their respective movements is short. From a trajectory of molecular dynamics, MONETA calculates for every pair of residues [43] (*i*) their *commute time* and (*ii*) the percentage *INT(a, b)* of conformations in which a non-covalent interaction exists between these two residues. Two residues communicate efficiently if their *commute time* is lower than a chosen value $CT_{threshold}$. A non-covalent interaction is considered as stable for *INT(a, b)* greater than a chosen threshold value $INT_{threshold}$. MONETA generates *CPs* iteratively according to the following algorithm:

- Start from one residue

- Create as many pathways {r_1, r_2 ...r_n} as residue *i*'s neighbors
- Grow each path{r_1, r_2 ...r_n} iteratively such that (*i*) r_i and r_{i+1} are not covalently bonded, (*ii*) there exists a stable non-bonded interaction between r_i and r_{i+1} , and (*iii*) r_{i+1} communicates efficiently with every residue in {r_1, r_2 ...r_i}.

The way *communication pathways* are grown ensures that any two adjacent residues are connected by non-covalent interactions, that all residues in a same path are directly or indirectly non-covalently connected and that every residue in the *CP* is reachable from any other point in a short *commute time*.

2. 3. Graphical implementation

Two-dimensional graphs: GEPHI [33] is an analytical tool designed for representation of interacting identities as a bi-dimensional graphic network. MONETA produces protein network graph (Figure 2) where all residues of the protein are denoted by as many nodes linked by edges representing connections of different natures: peptide links, direct connections (the two residues are consecutive in the same paths) and indirect connections (there exists at last one path linking the two residues). The global shape of the graph is constructed according to the inter-residues connections, so as to reflect the global communication propensity of the protein.

Three-dimensional visualisation: A PyMOL [32] module integrated in MONETA permits to visualize the communication pathways between the residues in the 3D structure of a protein (Figure 2). This representation gives access to a detailed vision of the communication pathways at the atomic level.

2.4. How does MONETA work?

The package distribution includes a step-by-step tutorial with description of input files, output files and shell scripts. The scripts can easily be modified to process user-provided input data. In brief, the first step of MONETA running requires a trajectory file obtained by MD simulations and the associated topology file. It generates all the prior data necessary for computing the *IDSs* and *CPs*:

- The structure onto which the MONETA representation will be drawn is by default chosen as the average structure of the protein in the MD simulation.

- Non-bonded interactions, *i.e.,* hydrogen bonds and hydrophobic contacts, are monitored along the trajectory and recorded using the program HBPLUS. MONETA runs HBPLUS on each conformation of the ensemble and computes the occurrence of non-covalently bonded interaction for each residues pair.
- The mean commute time of each residue pair over the MD simulation is computed.

  The *g_mdmat* or *g_mdmat_d* program of GROMACS is run by MONETA to compute, for each p*air of* resi*dues, the* minimal inter-residue distance averaged on the conformational ensemble.

A PCA is performed by use of the *ptraj* module of AMBER Tools to comp*ute the c*ovariance matrix, the PCA eigenvalues and eigenvectors.

To compute the data listed above with MONETA, the values of the required parameters listed below may be chosen by the user or computed automatically:

- The threshold for non-covalently bonded interaction occupancy ($INT_{threshold}$) beyond which a bond will be considered is 0.5 per default (*i.e.,* the interaction is present 50% of the time) but tunable at wish.
- The threshold for communication propensity ($CP_{threshold}$) – the inter-residue commute time below which the communication between these residues is considered as efficient – can be either arbitrary chosen or computed as a chosen percentile of all the values of communication propensity matrix.
- MONETA computes communications resulting from the 3D structure of a protein, rather than from its sequence. Consequently, a parameter sets the minimal distance along the sequence between two residues adjacent in a path. A value of 1 (residue), which turns off the «communications» mainly sustained by peptidic links, is recommended.
- The number of PCA modes to be considered in LFA is determined by setting a minimal part of the total atomic fluctuations of Cα- to be covered by these modes.
- The user can choose to remove some protein residues from the analysis (typically the strongly fluctuating extremities).

In the second step MONETA computes the *IDSs* and all the *CPs* within the protein. Then, MONETA computes for each residues pair {i,j} the number of paths containing the two residues and identifies the smallest paths between the two residues. The output files may be used for display of communication in the protein within the GEPHI [33] and PyMOL [32] modules.

PyMOL is automatically launched after *IDSs* and *CPs* computation. The structure file is loaded and represented as a green cartoon. *Independent dynamic segments* are loaded as PyMOL selections and highlighted in blue (tunable for each *IDS* in PyMOL selection column). Two functions, *find_paths* and *find_path2*, for visualization of *communication pathways* are loaded. Using these functions, the user can display all the *CPs* generated from a specified residue of the protein (*find_paths*), or the *CPs* between two specified residues (*find_path2*). A new PyMOL object is generated for each computed path and displayed as chains of small spheres connected together by lines (Figures 2, 4-6). Within a given path, each residue member is illustrated by up to three spheres corresponding to: the atom closest to the preceding residue in the path, the Cα, the atom closest to the following residue in the path. The initial residue is highlighted by a bigger sphere centered on its Cα.

GEPHI may be launched by the user to read the GEPHI outputs computed by MONETA and to create a 2D-representation of the communication network of the protein (Figures 2, 4 and 6). GEPHI uses two types of elements: nodes and edges. Nodes are auto-closing beacons containing an identification number (ID) and a label. The links have necessarily one ID, one label, a source (node of departure) and a target (node of arrival). To generate the GEXF files, MONETA generates automatically (i) the list of nodes from the topological file of the protein and (ii) the list of inter-residue links from both the topological file (peptidic links) and the files containing the communication pathways. Residues (nodes) are represented as circles, and connections between residues (edges) are represented by more or less thin lines. By default, large edges represent direct connection of two residues adjacent in a path and thin edges represent either a peptide link or a connection between two residues that are in the same path but not adjacent. MONETA computes the communication propensity of each residue as the number of other residues with which it communicates, and reports these data on the 2D- and/or 3D representation by scaled colors. These graphical parameters can be tuned at wish. The Force Atlas layout provided by GEPHI allows drawing the global shape of the protein network (Figure 3).

3. EXPLORING and PREDICTING ALLOSTERIC COMMUNICATION in RTKs and in its SIGNALLING PROTEIN STAT5

Receptor tyrosine kinases (RTKs) are cell-surface transmembrane receptors that possess a tightly regulated tyrosine kinase activity within their cytoplasmic domain [44]. RTKs are crucial components in signal transduction pathways involved in cellular differentiation, growth, survival and metabolism (Supplementary Information, Scheme S1). The crystal structures of their cytoplasmic region in activated forms are very similar [45], however crystal structures of inactive forms have revealed a remarkable variability in the kinase domain, which allows distinct activation mechanisms [46]. The description of the communication pathways mediating the allosteric regulation of the receptors activity is a challenging task. Furthermore, the deregulation of RTKs activity, mainly caused by mutations, is associated with various forms of cancers, neuronal disorders and inflammatory diseases. The motivation to understand disease processes and discover possible therapies, have driven our study of the type III RTK subfamily, that includes KIT, CSF-1R, PDGFR-α, PDGFR-β and FLT3 [47, 48] and of the proteins STAT5 involved in cell signaling pathways, downstream of various oncogenic proteins, particularly, of KIT [49].

Type III RTKs share a common architecture that includes an extracellular domain to which ligands bind, a transmembrane helix, an autoinhibitory juxtamembrane region (JMR) and a cytoplasmic region composed of the proximal and distal kinase domains (Supplementary Information**,** Scheme S1**,** A). Kinase activation, initiated through binding of its cognate ligand, stabilizes a dimeric form of the receptor which induces extensive conformational rearrangements involving the principal regulating fragments, A-loop and JMR [45, 50] and facilitates trans-autophosphorylation of tyrosine residues. These processes simulate tyrosine activity and creation of binding sites for downstream signaling proteins (Supplementary Information, Scheme S1, B).

Several gain-of-function point mutations induce tyrosine kinase constitutive (ligand-independent) activation and/or resistance to the tyrosine kinase inhibitors. For example, the equivalent mutations, D816V in KIT and D802V in CSF-1R, induce similar resistance effect, although only D816V confers a proliferative signal in KIT. The KIT mutations positioned in the JMR (V560G) or at the proximity of the Cα-helix (S628N) are highly oncogenic and provoke violent forms of cancers [51-52]. Cancer-related mutations frequently observed as a replacement by different amino acid residues (D816V/H/N/Y/E/I or V560G/D) have divergent impact on the severity of disease or drug resistance. Another issue consists in the role of the allosteric communication between JMR and A-loop on RTKs signaling. We proved that disruption of this communication in KIT mutant provokes a structural reorganization in the JMR, distant by more than 15 Å from the point mutation [53]. This important structural reorganization evidenced as a folding of the β-sheet of the JMR in KIT mutant should induce a distinct adaptation of the phosphotyrosine-based sites in the JMR affecting downstream signaling, which might not be the case in CSF-1R mutant. This hypothesis is consistent with the reported direct KIT – STAT5 interaction in the context of KIT D816V mutant (Supplementary Information, Scheme S1, D) that does not occur with the native KIT [49]. Consequently, description of the mutation-induced effects at the atomic level and discrimination between impacts of different mutations will contribute significantly to the understanding of mechanisms leading to deregulation of functional activation and signaling of RTKs.

We used MONETA to describe the allosteric regulation in the native receptors KIT and CSF-1R and in their numerous clinically observed mutants [29, 54-55].

First, the general landscape of *communication pathways* and the mapping of communication efficiency of residues, illustrated as 2D- and 3D-dimensional graphs (Figure 4) put in evidence the similarities between the two proteins and their particular features. Second, detailed inspection of *communication pathways* allowed us to establish the allosteric communication between (*i*) the JMR and the A-loop and (*ii*) the JMR and the Cα-helix, the major regulatory fragments in RTKs. Third, we established that all studied mutations (D816V/H/N/Y, V560G/D, S628N in KIT and D802V in CSF-1R) provoke (*a*) a diminishing of communication network in the mutants respectively to the native proteins, (*b*) the disruption of allosteric communication between essential regulatory fragments of the receptors (Figure 5), (*c*) a differential impact of the equivalent mutations on the communication in two homologous RTKs. Particularly, the differences between communication in KIT D816V and in CSF-1R D802V mutants indicate a divergent role of the equivalent mutations, which could explain why D802V mutation is not naturally found in cancer.

In addition, MONETA guided *in silico* mutagenesis as a way to restore the allosteric communication destroyed by oncogenic mutation [29].

MONETA was also applied on STAT5 proteins (STAT5a and STAT5b) and allowed to identify *communication pathways* connecting two spacially distant sites (>100 Å) of the protein in monomeric state (Figure 6). Phosphorylation of a specific tyrosine in the SH2 receptor binding domain of STAT5 controls allosteric regulation across this signaling protein mainly through the long α-helices of a Coiled-Coil Domain (CCD) acting as communication fibers. The different *communication pathways* within the Src-Homology domain 2 (SH2) and the Linker Domain (LD) in the unphosphorylated and the phosphorylated STAT5 proteins revealed the peculiar allosteric regulation in the activated and non-activated protein. Moreover, the *communication paths* in STAT5a and STAT5b evidenced a different role of phosphorylation of the specific tyrosine residue in these very similar proteins. It has been reported that limited structural dissimilarities between the STAT5a and STAT5b undoubtedly influence gene regulation [56], and the sequence-dependent asymmetries in STAT5 communication revealed by MONETA can be an underlying factor. Differences in a phosphorylation effect on the STAT5a and STAT5b communication pathways are other potential sources of functional disparity and the signal amplitude, frequency or duration which can also be significant [57].

4. CONCLUSIONS

MONETA is a freely available, easy-to-use and fully documented integrated package for analysis of the allosterically regulated proteins. The MONETA modules extract data from all-atom molecular simulations to detect, annotate and interpret local motions of functional relevance. It identifies dynamically coupled residues, *independent dynamics segments*, and builds a modular network representation composed of *IDSs* linked together by chains of residues, *communication pathways.* Such dual representation based on dynamics correlations and topological descriptors depicts signal propagation between allosterically coupled distant protein sites. This approach is an essential step toward the understanding of protein allosteric regulation and function.

The examples of a comprehensive analysis of communication in RTKs (KIT and CSF-1R) provided here put in evidence the well-established allosteric communication in the native proteins between the major regulating regions, A-loop, JMR and C-helix, that was disrupted by the point mutation localized in A-loop. A difference of communication in RTKs possessing equivalent mutations permits to distinguish between oncogenic mutation in KIT and resistance mutation in CSF-1R. The influence of phosphorylation on allosteric communication route established in STAT5 open a door for study of KIT-STAT5 signaling. An accurate description of the effector-induced (effector is a point mutation or phosphorylation) effects in terms of communication provides a solid basis for further guiding the discovery of next-generation allosteric drugs, capable to control protein functions.

A “restoring” second mutation in KIT mutant, predicted by MONETA, allowed modulating the communication between two remote principally regulatory segments, the A-loop and the JMR. The strong correlation between such communication and the structural and dynamical features of the receptors was established, illustrating an accurate predictive ability of MONETA.

Graphical tools, PyMOL and GEPHI, integrated in MONETA offer interactive 2D and 3D representations of communication at different scales – from a global landscape describing the inter-domains communication to inter-residues and inter-atomic pathways – to fully characterize the allosteric effects propagation in proteins. The next step of MONETA application will be the description of downstream signaling proteins activation.

MONETA addresses the need for automated functional analysis of allosteric communication emerging from molecular dynamics simulations. MONETA provides a solid basis for further guiding the discovery of next-generation allosteric drugs, with the aim to modulate or control protein function. The source code of MONETA is available by a simple request. Contacts: *Luba.Tchertanov@lbpa.ens-cachan.fr*.

Author Contributions

Conceived and designed the experiments: LT. Performed the experiments: AA FL EL. Analyzed the data: AA ICB FL YG EL. Wrote paper: LT ICB

Acknowledgements

This work was supported in part by funding from Ecole Normale Supérieure de Cachan, Institut Farman (ENS Cachan), Centre National de la Recherche Scientifique (CNRS), Laboratory of Excellence in Research on Medication and Innovative Therapeutics (LERMIT) and the French State Agency for Innovation (OSEO).

Figure legends

Figure 1. Schematic representation of the Modular NETwork Analysis (MONETA) general principle. A modular network representation composed of clusters of residues and chains of residues is built from the dynamical correlations and topology calculated from a protein conformational ensemble. In MONETA, residue clusters or modules are delineated as *independent dynamic segments* (*IDSs*) as they represent the most striking features of the protein local dynamics. Chains of individual residues are designated as *communication pathways* (*CPs*) as they represent well-defined connectivity pathways along which interactions can be mediated at long distances in the protein. Information is propagated through *IDSs* via the modification of the local atomic fluctuations and through CPs via well-defined interactions. The highly connected residues, at the junction of many pathways, can be considered as "*hubs*" in the protein network.

Figure 2. Overview of the major analysis steps in MONETA workflow. Each step of MONETA procedure is delimited by an icon. The required inputs, parameters, outputs and scripts are identified by colors: initial mandatory inputs in purple, outputs in blue, MONETA computation steps in green, software and program in grey. Step 3 is illustrated by 2D graph of communication landscape in KIT (**a**) and by 3D representation of communication pathway in STAT5 (**b**). 2D and 3D graphs drawn with GEPHI and PyMOL modules incorporated in MONETA.

Figure 3. Example of specific representation of inter-residues connection. Residues 9, 84, 25 and 13 represented by circles belong to a unique *communication pathway* and are linked in green (connected by at least one path or indirect connection) or in blue (adjacent in a path or direct connection). Residues 9 and 10 are linked in orange (peptide bond), but residue 10 does not belong to the path.

Figure 4. **Communication pathways of cytoplasmic region in the native receptors KIT (top) and CSF-1R (bottom).** *Left panel*: 2D graphs of a global topology of the inter-residues communications. Residues are represented by points, c*ommunication pathways* are depicted by bold lines and two connected residues by a thin line. Residues are coloured from blue through green and yellow to red according to their communication efficiency, estimated as the number of residues to which they are connected by at least one *CP*. *Middle panel*: Large-scale view of the *CPs* in KIT and CSF-1R zoomed on *CPs* in the JMR, the A-loop and C-loop. *Right panel*: 3D structural mapping of the inter-residues communication in KIT and CSF-1R. For each protein, the average MD conformation is presented as cartoon. The key structural fragments of receptors are highlighted in different colors: the A-loop is in red, the JMR is in blue, the Cα-helix is in magenta, the C- and P-loop are in green and yellow respectively. Labels for the different regions are indicated on the top panel. *Communication pathways* between residues atoms (circles) are depicted by coloured lines: *CPs* formed by the A-loop residues in orange; by the JMR residues in magenta. 2D and 3D graphs are drawn with GEPHI and PyMOL modules incorporated in MONETA.

Figure 5. **Characteristic features of the *communication pathways* generated in CSF-1R cytoplasmic region.** 3D structural mapping of the inter-residues communication in the native (*left*) and mutated (*right*) receptor. For each protein, the average MD conformation is represented as a cartoon. The key structural fragments of receptors are highlighted in different colors: the A-loop is in red, the JMR is in orange, the Cα-helix is in cyan, the C- and P-loop is in green and blue respectively. The c*ommunication pathways* between residues atoms (circles) are depicted by coloured lines: *CPs* formed by the C-helix residues in cyan; by the JMR residues in grey. 2D and 3D graphs are drawn with GEPHI and PyMOL modules incorporated in MONETA.

Figure 6. **Interaction network and modular network representation in the unphosphorylated (top) and the phosphorylated (bottom) STAT5.** *Left panels*: 2D graphs of a global topology the inter-residues communications. Residues are represented by points, c*ommunication pathways* are depicted by bold lines and two connected residues by a thin line. Residues are coloured from blue through green and yellow to red according to their communication efficiency, estimated as the number of residues to which they are connected by at least one *CP*. *Right panels*: 3D structural mapping of the inter-residues communication in STAT5. For each protein (unphosphorylated and phosphorylated), the average MD conformation is represented as a cartoon. The *independent dynamic fragments* are highlighted in different colors. *Communication pathways* between residues atoms (circles) are depicted by magenta. 2D and 3D graphs are drawn with GEPHI and PyMOL modules incorporated in MONETA. Large-scale view of the *CPs* in STAT5 zoomed on pathway within LD and SH2 domains indicating significant differences. Labels for the different regions of the proteins are indicated on the top panel.

Supplementary Information

Scheme S1. **Structural organization and possible signaling pathways of receptors tyrosine kinases (RTKs) (*e.g.*, KIT)**. (**A**) RTKs of type III comprise an extracellular ligand binding domain, a single transmembrane helix, a juxtamembrane region (JMR), a conserved kinase domains (proximal and distal) linked by a hinge and a carboxy-terminal tail (C-terminal). Phorphorylation sites are shown. (**B**) Activation of KIT induced by binding of SCF-ligand leads to dimerization, phosphorylation of specific tyrosine residues, and recruitment of several proteins at the intracellular portion of the receptors. Several proteins (*e.g.*, the cytokine receptor-associated Janus kinase, JAK2) bind directly to the receptor, whereas Ras/Raf pathways and AKT pathway need several specific adaptor molecules. JAK2 converts the latent monomeric form of the STAT molecules to the activated dimeric form through tyrosine phosphorylation. The dimers bind to specific DNA response elements and are able to induce transcription. All possible pathways result in nuclear activation of genes regulating cell growth, survival and mast cell activation. (**C**) Superimposed crystallographic structures of the cytoplasmic region of KIT and CSF-1R receptors in the inactive form show several differences. The proteins are presented as cartoon, CSF-1R is in light blue and KIT is in light grey. The key structural fragments of receptors in the inactive and the active conformations are highlighted in color. The JMR is in yellow and in orange ; the A-loop is in red and magenta; the Cα-helix is in cyan and blue, in KIT and CSF-1R respectively. (**D**) The constitutively activated receptor KIT stabilized by oncogenic point mutation (red spots with black contour), prompts alternative signaling routes, either through FES or by direct interaction with STAT5. These pathways result in nuclear activation of genes related with cell growth, survival and mast cell activation. Scheme was adapted from M. Arock's personal communication.

References

1. Nussinov, R. The spatial structure of cell signaling systems. *Phys. Biol.* **2013,** *10*, 045004.
2. Lepoivre, C.; Bergon, A.; Lopez, F.; Perumal, N. B.; Nguyen, C.; Imbert, J.; Puthier, D. TranscriptomeBrowser 3.0: introducing a new compendium of molecular interactions and a new visualization tool for the study of gene regulatory networks. *Bmc Bioinformatics* **2012,** *13*, 19.
3. Perkins, J. R.; Lees, J.; Antunes-Martins, A.; Diboun, I.; McMahon, S. B.; Bennett, D. L.; Orengo, C. PainNetworks: a web-based resource for the visualisation of pain-related genes in the context of their network associations. *Pain* **2013,** *154*, 2586-12.
4. Goel, A.; Li, S. S.; Wilkins, M. R. Four-dimensional visualisation and analysis of protein-protein interaction networks. *Proteomics* **2011,** *11*, 2672-2682.
5. Barabasi, A. L.; Gulbahce, N.; Loscalzo, J. Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.* **2011,** *12*, 56-68.
6. Doncheva, N. T.; Assenov, Y.; Domingues, F. S.; Albrecht, M. Topological analysis and interactive visualization of biological networks and protein structures. *Nat. Protoc.* **2012,** *7*, 670-685.
7. Tang, S.; Liao, J. C.; Dunn, A. R.; Altman, R. B.; Spudich, J. A.; Schmidt, J. P. Predicting allosteric communication in myosin via a pathway of conserved residues. *J. Mol. Biol.* **2007,** *373*, 1361-1373.
8. Atilgan, A. R.; Turgut, D.; Atilgan, C. Screened nonbonded interactions in native proteins manipulate optimal paths for robust residue communication. *Biophys. J.* **2007,** *92*, 3052-3062.
9. Atilgan, A. R.; Akan, P.; Baysal, C. Small-world communication of residues and significance for protein dynamics. *Biophys. J.* **2004,** *86*, 85-91.
10. Gandhi, P. S.; Chen, Z.; Mathews, F. S.; Di, C. E. Structural identification of the pathway of long-range communication in an allosteric enzyme. *Proc. Natl. Acad. Sci. U. S A* **2008,** *105*, 1832-1837.
11. Goncearenco, A.; Mitternacht, S.; Yong, T.; Eisenhaber, B.; Eisenhaber, F.; Berezovsky, I. N. SPACER: Server for predicting allosteric communication and effects of regulation. *Nucleic Acids Res.* **2013,** *41*, W266-W272.
12. Pasi, M.; Tiberti, M.; Arrigoni, A.; Papaleo, E. xPyder: a PyMOL plugin to analyze coupled residues and their networks in protein structures. *J. Chem. Inf. Model.* **2012,** *52*, 1865-1874.
13. Ghosh, A.; Vishveshwara, S. A study of communication pathways in methionyl- tRNA synthetase by molecular dynamics simulations and structure network analysis. *Proc. Natl. Acad. Sci. U. S A* **2007,** *104*, 15711-15716.
14. Ghosh, A.; Sakaguchi, R.; Liu, C.; Vishveshwara, S.; Hou, Y. M. Allosteric communication in cysteinyl tRNA synthetase: a network of direct and indirect readout. *J. Biol. Chem.* **2011,** *286*, 37721-37731.
15. Dixit, A.; Verkhivker, G. M. Computational modeling of allosteric communication reveals organizing principles of mutation-induced signaling in ABL and EGFR kinases. *PLoS Comput. Biol.* **2011,** *7*, e1002179.
16. Zheng, W.; Doniach, S. A comparative study of motor-protein motions by using a simple elastic-network model. *Proc. Natl. Acad. Sci. U. S A* **2003,** *100*, 13253-13258.
17. Panjkovich, A.; Daura, X. PARS: a web server for the prediction of Protein Allosteric and Regulatory Sites. *Bioinformatics* **2014**.
18. Juanico, B.; Sanejouand, Y. H.; Piazza, F.; De Los, R. P. Discrete breathers in nonlinear network models of proteins. *Phys. Rev. Lett.* **2007,** *99*, 238104.
19. Luccioli, S.; Imparato, A.; Lepri, S.; Piazza, F.; Torcini, A. Discrete breathers in a realistic coarse-grained model of proteins. *Phys. Biol.* **2011,** *8*, 046008.
20. Piazza, F.; Sanejouand, Y. H. Discrete breathers in protein structures. *Phys. Biol.* **2008,** *5*, 026001.
21. Lockless, S. W.; Ranganathan, R. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* **1999,** *286*, 295-299.
22. Manning, G.; Whyte, D. B.; Martinez, R.; Hunter, T.; Sudarsanam, S. The protein kinase complement of the human genome. *Science* **2002,** *298*, 1912-+.
23. Casaletto, J. B.; McClatchey, A. I. Spatial regulation of receptor tyrosine kinases in development and cancer. *Nat. Rev. Cancer* **2012,** *12*, 387-400.
24. Monod, J.; Wyman, J.; Changeux, J. P. On the nature of allosteric transitions: a plausible model. *J. Mol. Biol.* **1965,** *12*, 88-118.
25. Koshland, D. E., Jr.; Nemethy, G.; Filmer, D. Comparison of experimental binding data and theoretical models in proteins containing subunits. *Biochemistry* **1966,** *5*, 365-385.
26. Perutz, M. F. Stereochemistry of cooperative effects in haemoglobin. *Nature* **1970,** *228*, 726-739.
27. Cui, Q.; Karplus, M. Allostery and cooperativity revisited. *Protein Sci.* **2008,** *17*, 1295-1307.
28. Tsai, C. J.; del, S. A.; Nussinov, R. Protein allostery, signal transmission and dynamics: a classification scheme of allosteric mechanisms. *Mol. Biosyst.* **2009,** *5*, 207-216.
29. Laine, E.; Auclair, C.; Tchertanov, L. Allosteric communication across the native and mutated KIT receptor

tyrosine kinase. *PLoS Comput. Biol.* **2012,** *8*, e1002661.

30. Allain, A.; Chauvot de Beauchêne, I.; Laine, E.; Tchertanov, L. MONETA - Modular Network Analysis. 2013. ENS Cachan. Computer Program.
31. R Development Core Team. R: A Languege and Environment for Statistical Computing. -. 2013. R. Foundation for Statistical Computing.Vienna, Austria. ISBN 3-900051-07-0.
32. DeLano, W. L. The PyMOL Molecular Grapics System. 2002. Computer Program
33. Bastian, M.; Heymann, S.; acomy, M. Gephi: an open source software for exploring and manipulating networks. 2009. International AAAI Conference on Weblogs and Social Media. Computer Program
34. Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. The Amber biomolecular simulation programs. *Journal of Computational Chemistry* **2005,** *26*, 1668-1688.
35. Pronk, S.; Pall, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D.; Hess, B.; Lindahl, E. GROMACS 4.5: A high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics.* **2013**.
36. Mcdonald, I. K.; Thornton, J. M. Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.* **1994,** *238*, 777-793.
37. Piazza, F.; Sanejouand, Y. H. Long-range energy transfer in proteins. *Phys. Biol.* **2009,** *6*, 046014.
38. Penev, P. S.; Atick, J. J. Local feture analysis: a general statistical theory for object representation. *Network: Computation in Neural systems* **1996,** *7*, 477-500.
39. Zhang, Z. Y.; Wriggers, W. Local feature analysis: A statistical theory for reproducible essential dynamics of large macromolecules. *Proteins-Structure Function and Bioinformatics* **2006,** *64*, 391-403.
40. Chennubhotla, C.; Yang, Z.; Bahar, I. Coupling between global dynamics and signal transduction pathways: a mechanism of allostery for chaperonin GroEL. *Mol. Biosyst.* **2008,** *4*, 287-292.
41. Morra, G.; Verkhivker, G.; Colombo, G. Modeling signal propagation mechanisms and ligand-based conformational dynamics of the Hsp90 molecular chaperone full-length dimer. *PLoS Comput. Biol.* **2009,** *5*, e1000323.
42. Chennubhotla, C.; Bahar, I. Signal propagation in proteins and relation to equilibrium fluctuations. *PLoS Comput. Biol.* **2007,** *3*, 1716-1726.
43. Abraham, M. J.; Gready, J. E. Optimization of parameters for molecular dynamics simulation using smooth particle-mesh Ewald in GROMACS 4.5. *J. Comput. Chem.* **2011,** *32*, 2031-2040.
44. Blume-Jensen, P.; Hunter, T. Oncogenic kinase signalling. *Nature* **2001,** *411*, 355-365.
45. Huse, M.; Kuriyan, J. The conformational plasticity of protein kinases. *Cell* **2002,** *109*, 275-282.
46. Lemmon, M. A.; Schlessinger, J. Cell Signaling by Receptor Tyrosine Kinases. *Cell* **2010,** *141*, 1117-1134.
47. Robinson, D. R.; Wu, Y. M.; Lin, S. F. The protein tyrosine kinase family of the human genome. *Oncogene* **2000,** *19*, 5548-5557.
48. Ullrich, A.; Schlessinger, J. Signal transduction by receptors with tyrosine kinase activity. *Cell* **1990,** *61*, 203-212.
49. Chaix, A.; Lopez, S.; Voisset, E.; Gros, L.; Dubreuil, P.; De, S. P. Mechanisms of STAT protein activation by oncogenic KIT mutants in neoplastic mast cells. *J. Biol. Chem.* **2011,** *286*, 5956-5966.
50. Nolen, B.; Taylor, S.; Ghosh, G. Regulation of protein kinases; controlling activity through activation segment conformation. *Mol. Cell* **2004,** *15*, 661-675.
51. Frost, M. J.; Ferrao, P. T.; Hughes, T. P.; Ashman, L. K. Juxtamembrane mutant V560GKit is more sensitive to Imatinib (STI571) compared with wild-type c-kit whereas the kinase domain mutant D816VKit is resistant. *Mol. Cancer Ther.* **2002,** *1*, 1115-1124.
52. Vita, M.; Tisserand, J.; Chauvot de Beauchêne, I.; Panel, N.; Tchertanov, L.; Agopian, J.; Mascam_Mancini, L.; Fouet, B.; Fournier, B.; Dubreuil, P.; Bertucci, F.; De Sepulveda, P. Characterization of S628N, a novel *KIT* mutation found in a metastatic melanoma. *JAMA Derm* **2014,** *submitted*.
53. Laine, E.; Chauvot, d. B., I; Perahia, D.; Auclair, C.; Tchertanov, L. Mutation D816V Alters the Internal Structure and Dynamics of c-KIT Receptor Cytoplasmic Region: Implications for Dimerization and Activation Mechanisms. *PLoS Comput. Biol.* **2011,** *7*, e1002068.
54. Chauvot de Beauchêne, I.; Allain, A.; Laine, E.; Dubreuil, P.; Tchertanov, L. Hotspot mutations in KIT receptor differentially modulate its allosteric regulation: impact on activation and drug sensitivity. *PLoS Comput. Biol.* **2014,** *under revision*.
55. Da Silva Figueiredo Celestino Gomes, P.; Panel, N.; Laine, E.; Pascutti, P. G.; Solary, E.; Tchertanov, L. Does mutation D802V of the CSF1 Receptor alternates the tyrosine kinase tertiary structure and allosteric communication? PLoS ONE. *Submitted* **2014**.
56. Grimley, P. M.; Dong, F.; Rui, H. Stat5a and Stat5b: fraternal twins of signal transduction and transcriptional activation. *Cytokine Growth Factor Rev.* **1999,** *10*, 131-157.
57. Paukku, K.; Silvennoinen, O. STATs as critical mediators of signal transduction and transcription: lessons learned from STAT5. *Cytokine Growth Factor Rev.* **2004,** *15*, 435-455.

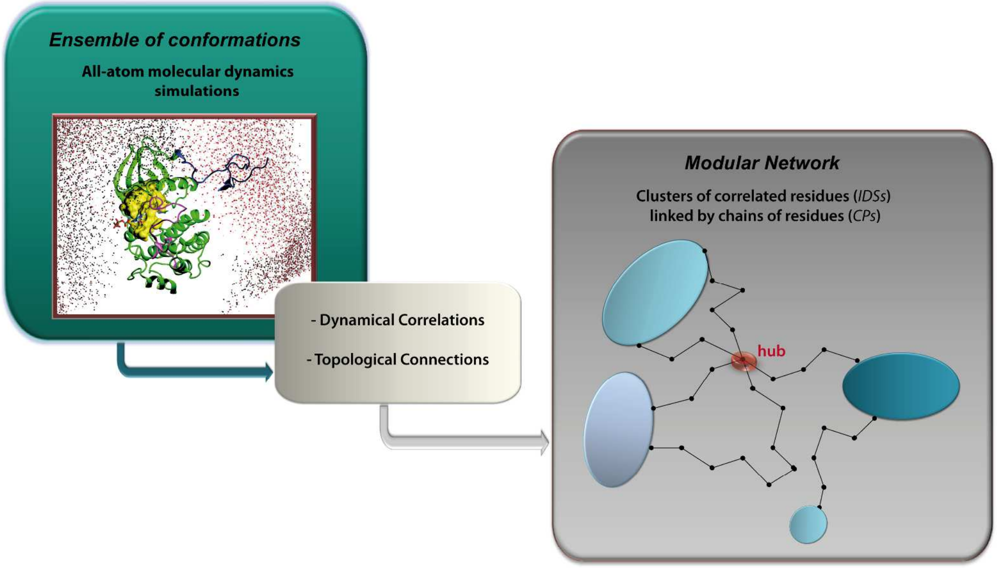

Ensemble of conformations
All-atom molecular dynamics simulations
- Dynamical Correlations
- Topological Connections
Modular Network
Clusters of correlated residues (*IDSs*) linked by chains of residues (*CPs*)
hub

Figure 1. Schematic representation of the Modular NETwork Analysis (MONETA) general principle. A modular network representation composed of clusters of residues and chains of residues is built from the dynamical correlations and topology calculated from a protein conformational ensemble. In MONETA, residue clusters or modules are delineated as independent dynamic segments (IDSs) as they represent the most striking features of the protein local dynamics. Chains of individual residues are designated as communication pathways (CPs) as they represent well-defined connectivity pathways along which interactions can be mediated at long distances in the protein. Information is propagated through IDSs via the modification of the local atomic fluctuations and through CPs via well-defined interactions. The highly connected residues, at the junction of many pathways, can be considered as “hubs” in the protein network.
258x148mm (300 x 300 DPI)

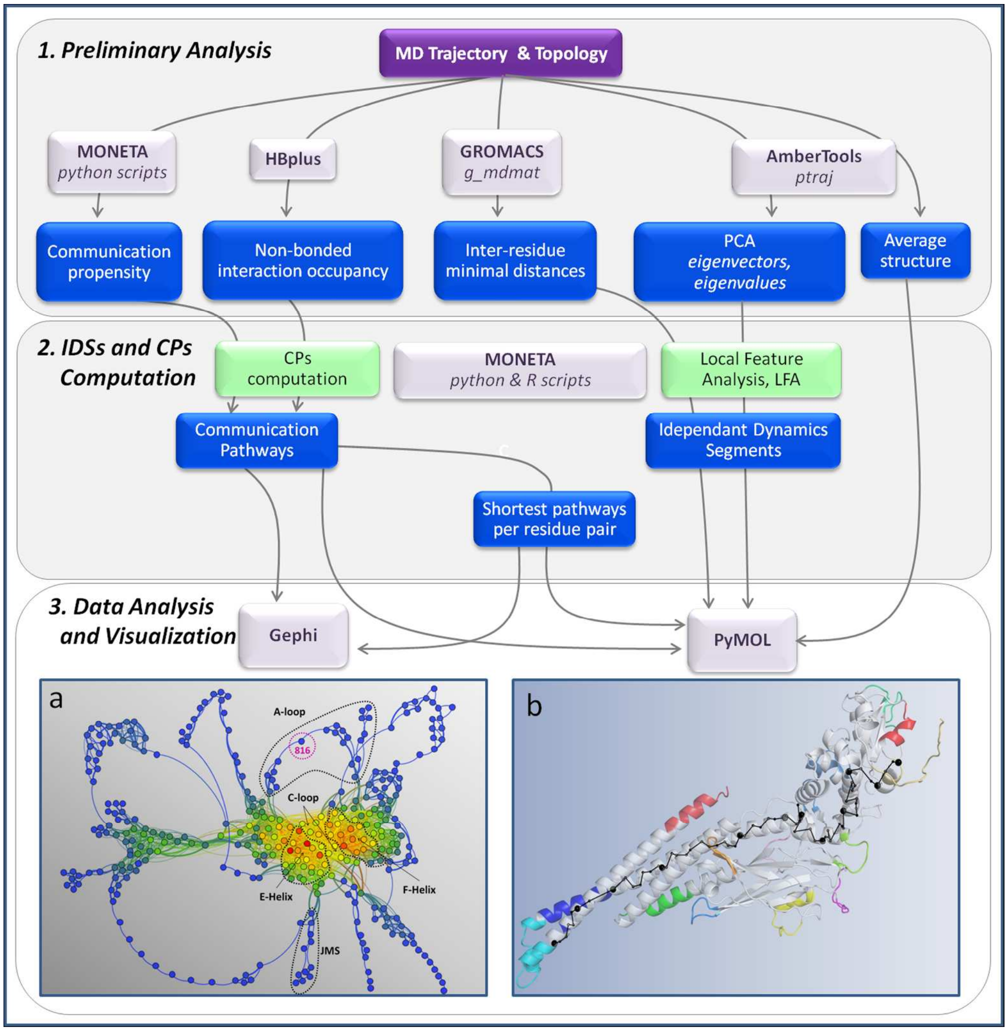


Figure 2. Overview of the major analysis steps in MONETA workflow. Each step of MONETA procedure is delimited by an icon. The required inputs, parameters, outputs and scripts are identified by colors: initial mandatory inputs in purple, outputs in blue, MONETA computation steps in green, software and program in grey. Step 3 is illustrated by 2D graph of communication landscape in KIT (a) and by 3D representation of communication pathway in STAT5 (b). 2D and 3D graphs drawn with GEPHI and PyMOL modules incorporated in MONETA.
199x204mm (300 x 300 DPI)

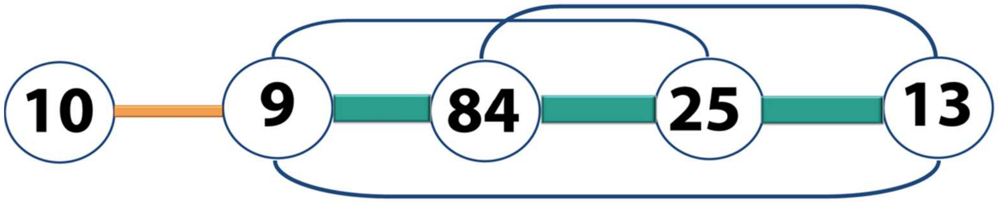


Figure 3. Example of specific representation of inter-residues connection. Residues 9, 84, 25 and 13 represented by circles belong to a unique communication pathway and are linked in green (connected by at least one path or indirect connection) or in blue (adjacent in a path or direct connection). Residues 9 and 10 are linked in orange (peptide bond), but residue 10 does not belong to the path.
99x20mm (300 x 300 DPI)

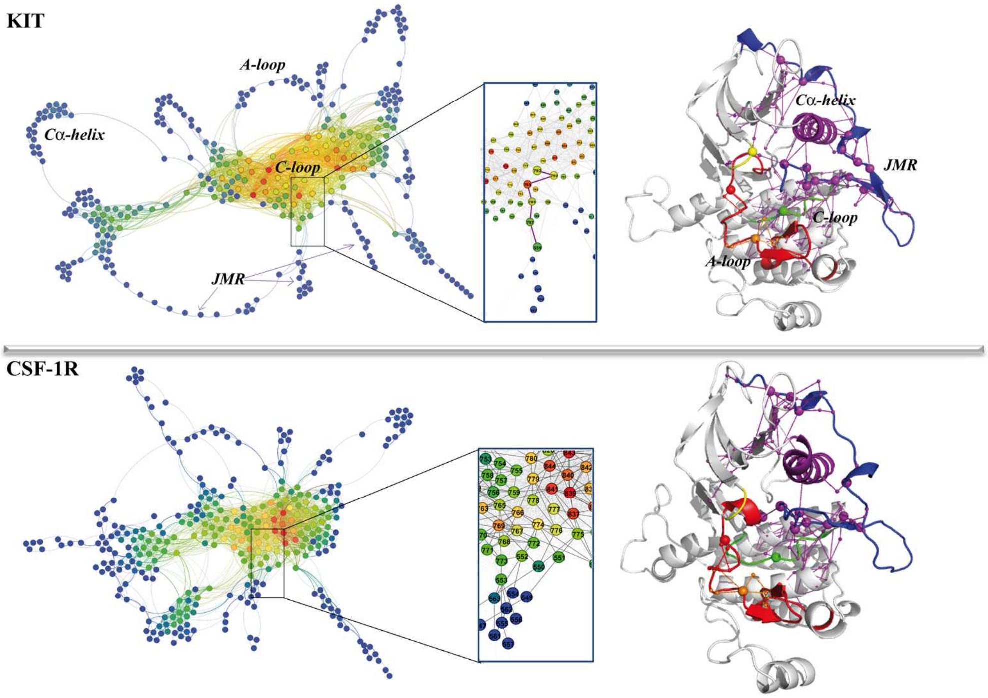


Figure 4. Communication pathways of cytoplasmic region in the native receptors KIT (top) and CSF-1R (bottom). Left panel: 2D graphs of a global topology of the inter-residues communications. Residues are represented by points, communication pathways are depicted by bold lines and two connected residues by a thin line. Residues are coloured from blue through green and yellow to red according to their communication efficiency, estimated as the number of residues to which they are connected by at least one CP. Middle panel: Large-scale view of the CPs in KIT and CSF-1R zoomed on CPs in the JMR, the A-loop and C-loop. Right panel: 3D structural mapping of the inter-residues communication in KIT and CSF-1R. For each protein, the average MD conformation is presented as cartoon. The key structural fragments of receptors are highlighted in different colors: the A-loop is in red, the JMR is in blue, the Cα-helix is in magenta, the C- and P-loop are in green and yellow respectively. Labels for the different regions are indicated on the top panel. Communication pathways between residues atoms (circles) are depicted by coloured lines: CPs formed by the A-loop residues in orange; by the JMR residues in magenta. 2D and 3D graphs are drawn with GEPHI and PyMOL modules incorporated in MONETA.

170x118mm (300 x 300 DPI)

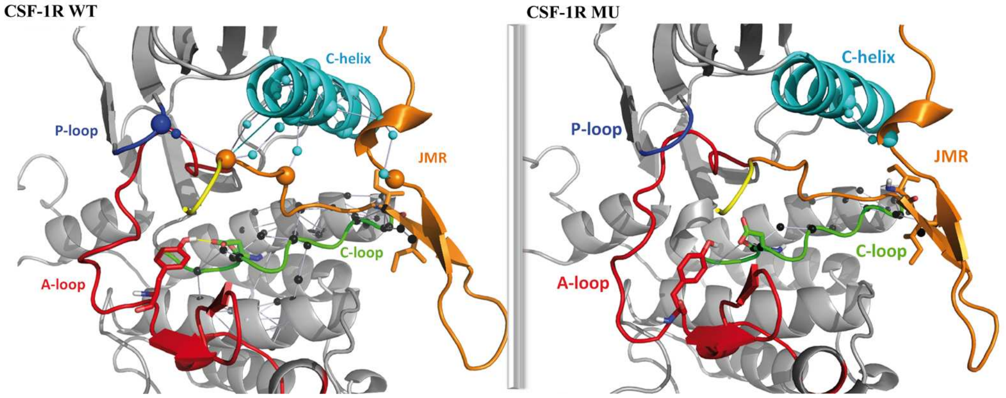


Figure 5. Characteristic features of the communication pathways generated in CSF-1R cytoplasmic region. 3D structural mapping of the inter-residues communication in the native (left) and mutated (right) receptor. For each protein, the average MD conformation is represented as a cartoon. The key structural fragments of receptors are highlighted in different colors: the A-loop is in red, the JMR is in orange, the Cα-helix is in cyan, the C- and P-loop is in green and blue respectively. The communication pathways between residues atoms (circles) are depicted by coloured lines: CPs formed by the C-helix residues in cyan; by the JMR residues in grey. 2D and 3D graphs are drawn with GEPHI and PyMOL modules incorporated in MONETA.
170x67mm (300 x 300 DPI)

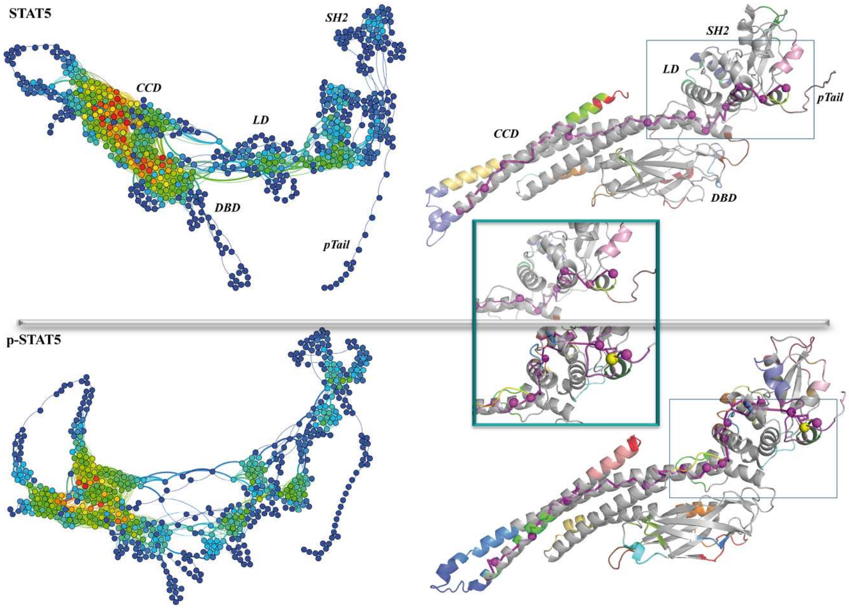


Figure 6. Interaction network and modular network representation in the unphosphorylated (top) and the phosphorylated (bottom) STAT5. Left panels: 2D graphs of a global topology the inter-residues communications. Residues are represented by points, communication pathways are depicted by bold lines and two connected residues by a thin line. Residues are coloured from blue through green and yellow to red according to their communication efficiency, estimated as the number of residues to which they are connected by at least one CP. Right panels: 3D structural mapping of the inter-residues communication in STAT5. For each protein (unphosphorylated and phosphorylated), the average MD conformation is represented as a cartoon. The independent dynamic fragments are highlighted in different colors. Communication pathways between residues atoms (circles) are depicted by magenta. 2D and 3D graphs are drawn with GEPHI and PyMOL modules incorporated in MONETA. Large-scale view of the CPs in STAT5 zoomed on pathway within LD and SH2 domains indicating significant differences. Labels for the different regions of the proteins are indicated on the top panel.
199x125mm (300 x 300 DPI)