Institut Farman FR 3311 : appel à projets AAP 2015

Proposition de projet Farman – Volet scientifique

ICAR : Identification de Chemins Allostériques à travers l'analyse de Réseaux

CMLA - LSV

Intitulé du projet : ICAR

Titre explicite : IDENTIFICATION DE CHEMINS ALLOSTERIQUES A TRAVERS L'ANALYSE DE RESEAUX

Version : Standard

Responsables scientifiques :

Trouvé Alain : tél. 01.47.40.59.18, email: trouve@cmla.ens-cachan.fr

Luba Tchertanov: tél. 01 47 40 76 62, email: Luba.Tchertanov@ens-cachan.fr

Stefan Haar: 01 47 40 75 67, email: stefan.haar@inria.fr

Durée du projet : 24 mois

Membres pressentis de l'équipe-projet :

Trouvé Alain (Pr), Bernard Chalmond (Pr), Jean Cazalis (stagiaire L3), Viannez Debavelaere (stagiaire L3), Chi-Tam Le (stagiaire L3) et Achille Samaran (stagiaire L3).

Luba Tchertanov (DR CNRS), Zoltan Palmai (Post Doc), Florent Langenfeld (Doctorant), Nolan Chatron (Doctorant), Simon Krief (stage M2)

Stefan Haar (DR INRIA), Thomas Chatain (MC, ENSC), Stefan Schwoon (MC-HdR-Chaire INRIA, ENSC), 1-2 doctorants à recruter.

Résumé du projet :

Le projet **ICAR** a pour objectif de définir les descripteurs valides et d'élaborer de nouvelles approches analytiques permettant l'interprétation des données de simulation de dynamique moléculaire des protéinées ciblées à l'échelle atomique. La représentation modulaire d'une protéine basée sur les propriétés dynamique et la description les *voies de communication* à travers la protéine constituent les éléments centraux dans la recherche visant à décoder l'allostérie – un phénomène fondamental en biologie – qui se manifeste par la transmission d'une information (signal) entre deux sites de la protéine spatialement distants. L'une des nouveautés de ce projet consistera en l'utilisation des méthodes informatiques pour l'analyse des processus d'événements discrets, tels que développés autours des réseaux de Pétri, afin d'identifier les composants dynamiques et d'appréhender leur interdépendances.

Description scientifique du projet

Introduction: Le projet **ICAR** s'intéresse à la description de la dynamique des protéines en vue de la compréhension de leurs fonctions à l'état physiopathologique. Cette description permet le développement d'une stratégie de modulation de celles-ci et la découverte de nouvelles pistes d'inhibition. Par conséquent, ce travail concerne en premier lieu les biologistes, les médecins et l'industrie pharmacologique. Le projet se base sur les compétences complémentaires des trois équipes-partenaires de l'institut FARMAN : mathématiciens (SIP, A. Trouvé, CMLA), informaticiens (MeXiCo, Stefan Haar, LSV) et les chercheurs en biologie computationnelle et structurale (BiMoDyM, L. Tchertanov, LBPA/CMLA). La dernière équipe (BiMoDyM) applique les méthodes de modélisation et simulation de dynamique moléculaire (DM) classique pour générer des données massives (*big data*) permettant la description dynamique des macromolécules. L'exploration des données générées, leurs représentations et l'analyse des résultats obtenus représente un défi majeur, surtout dans le contexte de l'étude des problématiques sophistiquées/complexes (par exemple, la résistance médicamenteuse ou la régulation allostérique des protéines). Par conséquent, l'application de méthodes mathématiques non-triviales et/ou le développement des nouvelles approches originales constituent une condition requise pour extraire une information complète et adéquate caractérisant le comportement des protéines et pour représenter cette information sous une forme optimale. Deux partenaires, BiMoDyM et SIP, travaillent ensemble depuis deux ans dans le contexte du projet FARMAN TopDyn (2013-2014). Les résultats obtenus montrent la nécessité d'une exploration plus profonde et plus large, ce qui constitue la base principale de notre projet actuel. Cette exploration étendue requiert l'implication d'un nouveau partenaire, MeXiCo, ayant une expertise absolument nécessaire pour élaborer ce nouveau projet ambitieux et complexe.

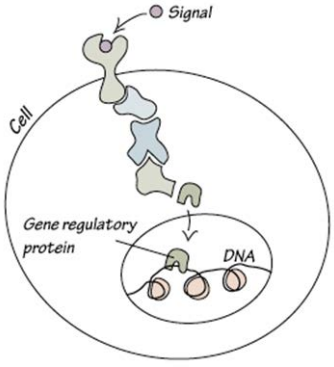

Figure 1. Schéma de la régulation allostérique au niveau cellulaire.

Contexte biologique et stratégie: Chaque processus biologique peut être décrit en termes physiques par des descripteurs associés à certaines formalisations mathématiques. Par exemple, la régulation allostérique des fonctions des protéines – un phénomène fondamental en biologie – résulte de la transmission d'une information (signal), induite par une perturbation locale (effecteur = ligand/substrat/inhibiteur/ mutation), entre deux sites de la protéine spatialement distants (**Figure 1**). Cette perturbation

mécanique conduit à une séquence de réarrangements conformationnels et une modification dynamique d'un ou plusieurs site(s) éloignés spatialement. De tels événements peuvent être décrits sous forme de transmission d'une information (communication) à longue portée à travers une protéine.

L'étude des phénomènes allostériques et des mécanismes de résistance nécessite la simulation d'un grand nombre de systèmes – différentes protéines dans divers états (phosphorylés ou non, liés à un effecteur ou non) et leurs nombreux mutants oncogéniques et/ou résistants. Ces simulations numériques sont très gourmandes en temps de calcul: il faut simuler de gros systèmes (le nombre de degrés de liberté dépasse plusieurs millions) et suffisamment longtemps (plusieurs millions de pas de temps pour les simulations de DM) pour avoir une représentation statistiquement précise du système à l'échelle macroscopique. De plus, les simulations à différentes échelles de temps, de la ns (10^{-9} s) à la µs (10^{-6} s) voire ms (10^{-1} s), caractérisent des mouvements différents (locaux, moyens, larges ou globaux) décrivant des fonctions variées (adaptation conformationnelle, transition allostérique ou activation de la protéine).

Cette complexité appelle le développement de méthodes de modélisation numérique avancées permettant de réduire le coût de calcul tout en assurant voire en améliorant la robustesse. Ces méthodes peuvent aborder un ou plusieurs des points suivants: aspects multi-physiques, multi-échelles, non linéarités géométriques et/ou hétérogénéité des comportements, non linéarités localisées**....**

Différentes conceptions de la communication intra-protéine: Il est communément admis que les voies de communication intra-protéine sont déterminées par la structure 3D et le réseau d'interactions liantes et non-liantes formé par les acides aminés de la protéine. Les effets allostériques sont souvent considérés comme la conséquence de chemins de communications intrinsèques entre domaines, c'est-à-dire préexistants à la perturbation transmise [1]. Un couplage structural *via* des chemins rigides entre sites effecteurs et sites affectés dans des protéines allostériquement régulées est mis en évidence [2]. La communication intra-protéine a aussi été décrite d'un point de vue thermodynamique, comme la conséquence de liens énergétiques plus ou moins directs entre des sites distants [3]. Il y aurait donc plusieurs types de mécanismes allostériques, basés sur différents liens de causalité entre repliement/rigidité et flexibilité des sites effecteurs et affectés et des segments les reliant. Ces visions de l'allostérie comme un phénomène global à l'échelle de la protéine ne contredisent pas l'idée de transmission de proche en proche de perturbations locales. Des communications allostériques par la transmission de fluctuations de chaînes latérales ont été mises en évidence en 2011 par dynamique de Monte Carlo [4]. Cette coexistence de phénomènes globaux et locaux est une autre illustration de la diversité des mécanismes de communication intra-protéine.

Résultats antérieurs: **MOdular NETwork Analysis - MONETA** : Afin de décrire les phénomènes allostériques au niveau atomique, L. Tchertanov et ses collaborateurs ont mis en place une approche originale, MOdular NETwork Analysis (MONETA), conçue pour localiser la propagation d'une perturbation au sein d'une protéine [5-7] (**Figure 2**).

MONETA utilise des données statistiques sur des ensembles conformationnels issus de simulations de Dynamique Moléculaire (DM) « tout atome » relatives d'une part à la topologie de la protéine (interactions inter-résidus ou *Communication Pathways, CPs*) et d'autre part aux corrélations dynamiques inter-résidus *(Independent Dynamic Segments, IDSs*).

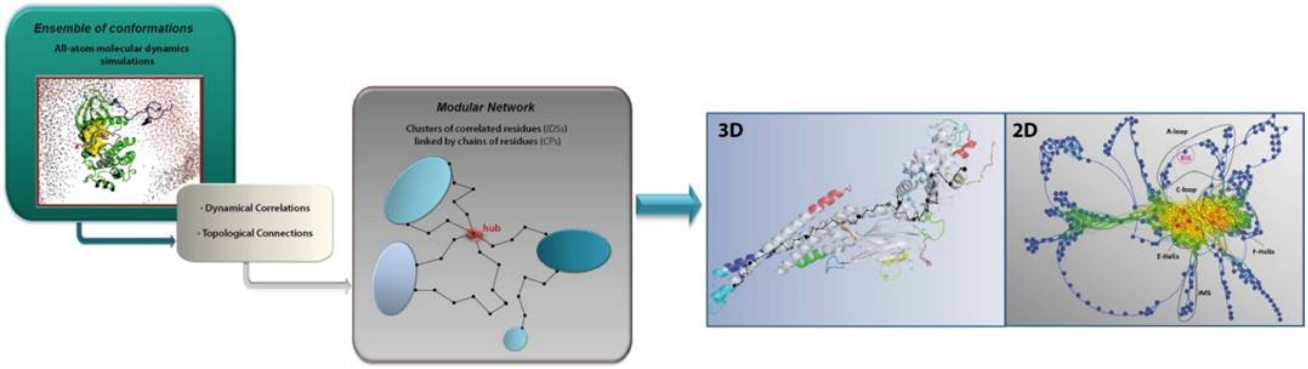

Figure 2. MONETA : Représentation en réseau modulaire d'une protéine et Visualisation des communications inter-résidus en 3D and 2D.

Avec cet outil, la régulation allostérique de certaines protéines impliquées dans la transduction du signal cellulaire, les récepteurs tyrosine kinases et leurs nombreux mutants recensés par les cliniciens, ont été étudiés par BiMoDyM et en collaboration avec SIP (A. Trouvé) [5,7-10]. La méthode a été utilisée par d'autres chercheurs, *i.e.*, dans [11-12], et commentée par les experts du domaine [13-14].

Cependant, au vu de l'expériences acquises sur l'analyse de différentes protéines [15-16], il est clair que la complexité des phénomènes en jeu demande d'intensifier le travail de clarification et formalisation de l'analyse des données issues des simulations de dynamique moléculaire du côté des mathématiques, des statistiques et de l'informatique (qui est prise ici au sens de l'analyse des système d'informations). Par exemple et de façon surprenante, peu de choses sont connues sur les propriétés statistiques des trajectoires générées en terme de variabilité et stabilité par rapport aux conditions initiales et surtout sur l'impact des états métastables visités aux différentes échelles de temps sur les descripteurs utilisés dans l'analyse des dynamiques conformationnelles, ce qui rend problématique l'analyse de l'impact sur la dynamique d'un effecteur (fixation d'un ligand, apparition d'une mutation…).

Un premier résultat obtenu d'une approche plus formalisée a été la construction d'une méthode de détection des *IDS*s s'appuyant sur une modélisation plus explicite en amont des propriétés statistiques décrivant les objets dynamiques recherchés et permettant d'élaborer des algorithmes de détection adaptée. Ceci a permis le développement d'une nouvelle méthode, la PFD (**Principal Features Décomposition**) en collaboration avec A. Trouvé (SIP/CMLA) dans le cadre du projet FARMAN TopDyn 2013-2014. La méthode PFD, alternative à la méthode *sans modèle* LFA initialement intégrée dans MONETA, permet d'avoir une approche plus explicite au niveau de la modélisation des propriétés des briques de bases que sont les *IDSs*. La nouvelle méthode a été testée sur des données de simulation de DM de KIT (les résultats font l'objet d'une publication en cours de rédaction). La méthode PFD a été également appliquée pour analyser les trajectoires de 30 et 200 ns des protéines STAT5a et STAT5b et de leurs homologues phosphorylés [18]. L'article présentant la méthode PFD et sa validation par analyse des plusieurs protéines est en cours de rédaction. En parallèle, un certificat de protection des droits des auteurs sera déposé le printemps de 2015.

Nouveaux enjeux :

Le projet **ICAR** a pour objectif de définir les descripteurs et d'élaborer de nouvelles approches analytiques permettant la description des mécanismes de la régulation allostérique des protéines ciblées à l'échelle atomique. Il vise en particulier l'extension, au domaine de l'analyse des phénomènes

de régulation allostérique des protéines, du principe d'une modélisation plus formalisée en préalable au développement des outils d'analyse des données de simulation de dynamique moléculaire. Plus encore, il s'agit d'explorer l'apport des concepts et des idées venant de l'étude formelle des systèmes d'informations à la modélisation des phénomènes de communication dynamique qui sous-tend au niveau atomique le phénomène de régulation allostérique par une collaboration avec l'équipe INRIA MeXiCo du LSV. L'équipe MeXiCo du LSV travaille sur l'analyse formelle des systèmes concurrents à événements discrets, en se focalisant sur les modèles permettant d'exhiber les relations de causalité, tels les réseaux de Pétri et leur sémantique en ordre partiel. Plusieurs résultats des dernières années concernent la détection efficace de relations de corrélation indirectes entre processus (nommée *reveals*) [19-21,23], ainsi que l'identification d'attracteurs dans les réseaux de régulation génétique [24]. L'équipe a également utilisé des modèles formels d'évolution de graphe comme système dynamique [22]. Ces expertises et connaissances sont essentielles pour approfondir l'étude de dépendances indirectes en apportant des concepts et résultats utiles à la compréhension des *IDS*s, ainsi que concernant l'émergence et l'évolution de la communication intra-protéine.

Du point de vue des outils d'analyse de la dynamique, les travaux menés entre l'équipe SIP et BiMoDyn ont montré que les *IDSs* sont des entités statistiques élémentaires qui constituent des éléments terminaux dans une description hiérarchique de la dynamique dans une fenêtre temporelle. L'existence de mouvement globaux à l'échelle de la protéine engendre des variables latentes qui génèrent de la dépendance entre les *IDSs* qui doivent être vus comme seulement conditionnellement indépendants, connaissant l'existence de ces variables latentes. La représentation naturelle qui se dégage est celle d'un réseau bayésien structurant plusieurs niveaux de représentation (au minimum un niveau global à l'échelle de la protéine tout entière, un niveau intermédiaire correspondant à des versions dynamiques des domaines fonctionnels et un niveau bas correspondant aux *IDSs*). L'identifiabilité d'une telle décomposition est seulement partielle et posent de difficiles problèmes nécessitant de combiner à la fois les aspects modélisation en lien avec la biologie mais aussi le développement de nouvelles techniques d'estimation. La prise en compte de la géométrie de la molécule dans la décomposition de la dynamique est insuffisante dans l'approche actuelle et pourrait fournir des *a priori* pour la construction de l'analyse hiérarchique que nous comptons exploiter. Par exemple, les techniques récentes de décomposition de matrice en somme d'une matrice de rang faible et d'une matrice éparse par des méthodes d'optimisation convexe se prêtent à l'introduction de pénalisation géométrique sur la matrice de rang faible qui ouvre des pistes intéressantes.

Du point de vue plus spécifique de l'analyse des voies de communication entre des sites distants d'une protéine, de nombreuses questions au niveau conceptuel se posent : quel(s) type(s) de composants (atomes, résidus ou fragments structuraux) pouvons-nous considérer comme les « *graines-porteurs/conducteurs* » valables dans la description de la communication ? Quelle type d'interactions (liaisons hydrogènes, interaction ionique, van-der-Waals) contribuent à la communication ? Quelle est l'impact des éléments conservés et variables au niveau de la description séquence/structure/dynamique sur la communication ? La géométrie spatiale est également un élément clé pour la définition des *chemins de communications* (*CPs*) qui sont construits à partir de relations de proximité entre résidus qui sont dynamiquement stables. Cependant les techniques actuelles sont construites sur des statistiques calculées sur la mesure empirique générée par l'ensemble des configurations et n'analysent pas la dynamique comme une succession d'événements temporellement reliés. La recherche des *voies*

de communication dans les phénomènes allostériques dans les protéines ciblées pourrait sans doute être considérablement enrichie par la prise en compte *des phénomènes temporels comme l'établissement ou la destruction de liaison non-covalente entre résidus ouvrant ou fermant des chemins de circulation.* Une modélisation de cette dynamique de transition comme un processus markovien de saut par un modèle paramétrique simple fonction des inter-distances entre résidus et des mesures de corrélations canoniques calculé sur les données est prévu. Il est nécessaire d'introduire des outils conceptuels nouveaux pour utiliser ce type d'information et les techniques d'analyse venant de l'informatique de systèmes d'événements discrets en particulier des réseaux de Pétri capables d'appréhender des systèmes complexes seront explorées dans le projet.

Originalité du projet

L'originalité du projet vient de la réunion d'expertises très différentes permettant d'aborder l'analyse des données de simulations de dynamique moléculaire d'un point de vue large et d'assurer la prise en compte à la fois du contexte biologique sous-jacent mais aussi de la nécessité de la prise en compte nécessaire de modèles formels préalables à l'analyse synthétique des données. Par ailleurs, si la modélisation des voies de signalisation et de régulation intra- et inter-cellulaires ont déjà fait l'objet d'une formalisation assez avancée du point de vue mathématique et informatique avec l'introduction en particulier de modélisation par réseaux dynamiques et stochastiques, il n'y a pas d'équivalent au niveau de la protéine. L'analyse de la dynamique de la protéine du point de vue des systèmes à événements discrets est également très différente du point de vue qui prévaut actuellement dans la communauté où les trajectoires sont considérées au travers des mesures empiriques.

Valeur ajoutée des différents partenaires à la réalisation du projet

Le projet **ICAR** a été inspiré au cours d'échanges avec nos collègues cliniciens. Il est basé sur la connaissance complémentaire des partenaires du projet. Le rapprochement des compétences en biologie computationnelle et structurale (BiMoDyM), en statistiques et en géométrie (SIP) des partenaires du CMLA, et celle en analyse formel des systèmes à événements discret des partenaires de l'équipe MeXiCo au sein du projet **ICAR** permettra d'explorer de façon originale l'utilisation d'un large corpus de données de simulations de dynamique moléculaire obtenue par l'équipe de Luba Tchertanov. Le développement d'outils originaux adaptés aux questionnements des cliniciens/biologistes ne peut se faire que si une réelle collaboration sur les besoins, les attentes et le savoir-faire des uns et des autres se met en place. Le projet est basé sur une gamme de concepts et par conséquent, c'est un projet interdisciplinaire et exploratoire. Une collaboration au sein de l'institut Farman semble être le lieu approprié pour cela. Les résultats de ces travaux seront résumés et publiés dans des articles communs. Dans le projet seront associés plusieurs étudiants en License (co-encadrés au sein de équipes SIP et BiMoDyM), en Master et en Thèse des équipes-partenaires.

Références

1. del Sol A. et al., The origin of allosteric functional modulation: multiple pre-existing pathways. (**2009**). *Structure,* **17**(8): 1042-50.
2. Rader A. J. and S. M. Brown, Correlating allostery with rigidity. *Mol Biosyst,* 2011. **7**(2): p. 464-71.
3. Wrabl J. O. et al., The role of protein conformational fluctuations in allostery, function, and evolution. (**2011**). *Biophys Chem,* **159**(1): p. 129-41.
4. Dubay K. H. et al. (**2011**). Long-range intra-protein communication can be transmitted by correlated side-chain fluctuations alone. *PLoS Comput Biol,*. **7**(9): p. e1002168.
5. Laine, E., Auclair, C. and Tchertanov, L. (**2012**). Allosteric Communication across the Native and Mutated KIT Receptor Tyrosine Kinase. *PLoS Comput Biol*. 8(8): e1002661., 14 pages; doi:10.1371/journal.pcbi.1002661.
6. INTERDEPOSIT CERTIFICATION. (**2014**). Certificat délivré par Agence pour la Protection des Programmes. Inter Deposit Digital Number IDDN.FR.001.020012.000.S.P.2014.000.31235. Pour l'œuvre: MOdular NETwork Analysis (MONETA) version 2.0 en date du 31 juillet 2013. Les auteurs: Tchertanov L, Laine E., Allain A., Chauvot de Beauchêne I.
7. Allain A., Chauvot de Beauchêne I., Langenfeld F., Guarracino Y., Laine E., and Tchertanov L. (**2014**). Allosteric Pathway Identification through Network Analysis from Molecular Dynamics Simulations to Interactive 2D and 3D Graphs. *Faraday Disc.*, 169, 1-18. DOI:10.1039/C4FD00024B.
8. Da Silva Figueiredo Celestino Gomes P., Panel N., Laine E., Pascutti P. G., Solary E. and Tchertanov L. (**2014**). Differential effects of CSF-1R D802V and KIT D816V homologous mutations on receptor tertiary structure and allosteric communication. *PLoS ONE*. May 14;9(5):e97519. doi: 10.1371/journal.pone.0097519.
9. Vita M, Tisserand J C, Chauvot de Beauchêne I, Panel N, Tchertanov L, Mescam-Mancini L, Agopian J, Fouet B, Fournier B, Dubreuil P, Bertucci F, and De Sepulveda P. (**2014**). Characterization of S628N, a novel KIT mutation found in a metastatic melanoma. *JAMA Dermatology*, doi:10.1001/jamadermatol.2014.143.Published online October 2014.
10. Chauvot de Beauchêne I., Alain A., Panel N., Laine E., Trouvé A., Dubreuil P. and Tchertanov L. (**2014**). Oncogenic mutations of KIT receptor differentially modulate tyrosine kinase activity and drug susceptibility. *PLOS Comput. Biol*.10(7):e1003749. doi: 10.1371/journal.pcbi.1003749.
11. Palmai Z, Seifert C, Gräter F, Balog E. An allosteric signaling pathway of human 3-phosphoglycerate kinase from force distribution analysis. PLoS Comput Biol. 2014 Jan;10(1):e1003444. doi: 10.1371/journal.pcbi.1003444. Epub 2014 Jan 23.
12. Eren D, Alakent B. Frequency response of a protein to local conformational perturbations. PLoS Comput Biol. 2013;9(9):e1003238. doi: 10.1371/journal.pcbi.1003238. Epub 2013 Sep 26.
13. Tsai CJ, Nussinov R. A unified view of "how allostery works". PLoS Comput Biol. 2014 Feb 6;10(2):e1003394. doi: 10.1371/journal.pcbi.1003394. eCollection 2014 Feb.
14. Huang Z, Mou L, Shen Q, Lu S, Li C, Liu X, Wang G, Li S, Geng L, Liu Y, Wu J, Chen G, Zhang J. ASD v2.0: updated content and novel features focusing on allosteric regulation. Nucleic Acids Res. 2014 Jan;42(Database issue):D510-6. doi: 10.1093/nar/gkt1247. Epub 2013 Nov 28.
15. Arlet JB, Ribeil JA Guillem F, Negre O, Hazoume A, Marcion G, Beuzard Y, Dussiot M, Moura IC, Demarest S, Chauvot de Beauchêne I, Belaid-Choucair Z, Sevin M, Maciel TT, Auclair C, Leboulch P, Chrétien S, Tchertanov L, Baudin-Creuza V, Seigneuric R, Fontenay M, Garrido C, Hermine O, Courtois G. (**2014**). Heat shock protein 70 cytosolic sequestration by excess of free α-globin chains is a key mechanism of the ineffective erythropoiesis in β-thalassemia major patients. *Nature*.2014 Aug 24. doi: 10.1038/nature13614.
16. Gardie B., Couvé, S., Ladroue, C., Laine, E., Mathouk, K., Guégan, J., Gad, S., Lejeune, H. Lecomte, B., Pagès, J.-C., Collin, C., Lasne F., Bressac de Paillerets, B., Feunteun, J. Dessen, P., Lazar,V., Tchertanov, L., Mole, D., Kaelin, W., Ratcliffe, P., Richard, S. (**2014**). A comprehensive study of germline mutations in the VHL gene reveals the importance of precisely tuned dysregulation of the hypoxia pathway in oncogenesis. *Cancer Res*. 2014 Nov 15;74(22):6554-64. doi: 10.1158/0008-5472.CAN-14-1161.
17. Charon N. and Trouvé A. (**2013**). The varifold representation of non-oriented shapes for diffeomorphic registration. SIAM Journal of Imaging Science 6(4): 2547-2580
18. Langenfeld F., Guarracino Y., Arock M., Trouvé and Tchertanov L. (**2015**). How intrinsic molecular dynamics controls intramolecular communication in Signal Transducers and Activators of Transcription Factor STAT5. PLoS Comput Biol. Submitted 20 Jan. 2015
19. Th. Chatain and S. Haar. A Canonical Contraction for Safe Petri Nets. *In* Transactions on Petri Nets and

Other Models of Concurrency IX, LNCS 8910, pages 83-98. Springer, 2014
20. S. Haar, C. Kern and S. Schwoon. Computing the Reveals Relation in Occurrence Nets. Theoretical Computer Science 493, pages 66-79, 2013.
21. S. Balaguer, Th. Chatain and S. Haar. Building Occurrence Nets from Reveals Relations. Fundamenta Informaticae 123(3), pages 245-272, 2013.
22. P. Baldan, Th. Chatain, S. Haar and B. König. Unfolding-based Diagnosis of Systems with an Evolving Topology. Information and Computation 208(10), pages 1169-1192, 2010.
23. S. Haar. Types of Asynchronous Diagnosability and the Reveals-Relation in Occurrence Nets. IEEE Transactions on Automatic Control 55(10), pages 2310-2320, 2010.
24. Th. Chatain, S. Haar, L. Jezequel, L. Paulevé and S. Schwoon. Characterization of Reachable Attractors Using Petri Net Unfoldings. *In* CMSB'14, LNBI 8859, pages 129-142.

Publication du projet scientifique sur site web Farman : OUI