

Contractualisation vague D 2010-2013

Unité de recherche : dossier unique

Demande de reconnaissance

(Partie 1 : Projet scientifique)

Le dossier unique de demande de reconnaissance d'une unité de recherche a été remanié pour prendre en compte les nouvelles procédures d'évaluation (AERES).

Il constitue une plate-forme commune d'informations pertinentes pour l'évaluation des unités par cette agence et pour le pilotage par les établissements d'enseignement supérieur et de recherche, les organismes et l'Etat.

Les parties bilan et projet ont été séparées, ce qui permet de donner toute son importance au bilan dont le contenu a été étoffé, et de faciliter l'expertise des unités en recomposition (fusions, éclatements...).

1 – Projet scientifique pour la période 2010-2013

Le projet scientifique comprendra de 10 à 50 pages en fonction de la taille de l'unité et ne dépassera pas 4 Mo dans sa version électronique. Il pourra être rédigé en anglais ; dans ce cas, le résumé sera rédigé en français et en anglais.

Le projet doit permettre d'apprécier la pertinence des objectifs affichés par l'unité au regard de ses missions, de sa taille, de son organisation, et le cas échéant de sa capacité à avoir réalisé le projet précédent. Il précisera les dispositifs mis en œuvre pour atteindre ces objectifs, notamment :

- *la politique d'incitation à l'émergence de sujets innovants, à la prise de risque et aux sujets frontières, l'adaptation aux évolutions dans le contexte local, national, européen et international, la réflexion et la prospective sur les créneaux scientifiques porteurs, la vision de l'évolution de l'unité à 4 et 8 ans*
- *la politique d'animation de l'unité*
- *la politique de recrutement, l'analyse prospective à moyen et long terme des besoins et des compétences nécessaires aux mutations scientifiques et techniques identifiées, la politique de formation*
- *la politique d'investissement et sa cohérence avec le projet scientifique de l'unité, en précisant les principales acquisitions envisagées et les éventuels cofinancements prévus*
- *la politique de répartition des moyens, en particulier pour les unités finalisées, en fonction des objectifs fixés*
- *la capacité de l'unité à valoriser ses travaux de recherche*
- *son implication en matière de diffusion de l'information scientifique et technique*

L'unité dans son ensemble ou certaines de ses équipes peuvent porter un projet de recherche technologique et/ou de transfert. Dans ce cas, les objectifs visés en termes de ruptures et verrous technologiques seront précisés et les partenariats développés avec le tissu socio-économique seront décrits.

On précisera l'organigramme prévu de l'unité.

1.1 Projet collectif

La réunion dans les mêmes murs des deux équipes a constitué un tournant important dans la vie de l'unité. Si nous conservons la structure en équipes pour des raisons de visibilité, notre fonctionnement quotidien vise au contraire à mêler le plus possible nos travaux et nos expériences. Afin d'augmenter encore ces interactions, nous nous proposons de travailler sur un sujet largement fédérateur : l'inférence dans le cadre de modèles complexes.

1.1.1 Modèles complexes

De la biologie moléculaire à l'écologie, tous les domaines de la biologie moderne ont aujourd'hui recours à des modélisations complexes pour décrire le comportement et les interactions d'entités nombreuses, à différentes échelles, dans l'espace et le temps. Pour rendre compte de tels phénomènes, il est nécessaire de distinguer deux niveaux de structuration : le processus de recueil des données et le processus d'intérêt sous-jacent.

Ainsi les captures de thons effectuées chaque année sont les données à partir desquelles on essaie d'inférer le processus de renouvellement de stock. De même, les corrélations observées entre les niveaux d'expressions des gènes révèlent une structuration en classes fonctionnelles qu'on cherche à déterminer. Les modèles graphiques, hiérarchiques ou à variables latentes permettent de décrire ces deux niveaux de variabilité avec une grande souplesse. Leur représentation sous forme de graphe illustre leur caractère modulaire.

De plus, on peut souvent supposer que le processus latent génère des informations de natures différentes (volume de pêches et expériences de capture – marquage – recapture ou données de transcriptome et de protéome). La modélisation hiérarchique permet ainsi d'intégrer ces différentes mesures dans un même modèle.

Cette grande souplesse de modélisation est particulièrement souhaitable pour les applications biologiques intéressant l'UMR. Elle se paie cependant par une étape d'estimation plus difficile : les stratégies classiques d'inférence aussi bien bayésiennes que fréquentistes deviennent alors fréquemment impossibles à mener. Nous nous proposons donc de développer des méthodes d'estimation efficaces en suivant deux axes principaux.

Algorithmes stochastiques

De nombreuses techniques d'inférence requièrent de simuler un échantillon selon une loi particulière. Les méthodes MCMC constituent un exemple classique, de même que la version stochastique de l'algorithme EM (SEM). Dans de nombreux cas, ces méthodes aboutissent à des temps de calculs prohibitifs. Par exemple dans le cadre des modèles individus centrés (modèles de croissance d'arbre, cycle de vie des thons) où le comportement de la population est l'agrégation de comportements individuels, ces algorithmes sont inutilisables.

Nous souhaitons donc développer des compétences dans les algorithmes particuliers. L'idée centrale de ces algorithmes est de construire un échantillon de manière itérative en l'améliorant pas à pas. Ils dérivent du filtre de Kalman et peuvent être utilisés à la fois pour la simulation et pour l'estimation (particulièrement dans un cadre bayésien). Développer originellement pour le filtrage séquentiel, ces algorithmes n'ont été que peu utilisés dans un cadre d'estimation « statique ».

Nous souhaitons développer l'usage de ces algorithmes particulièrement dans le cadre des modèles en foresterie et un projet sur ce thème est déposé en collaboration avec l'INRIA et le CIRAD de Montpellier dans le cadre du *RTRA Agronomie et Développement Durable*.

Méthodes approchées

Une autre façon de mener l'inférence pour des modèles complexes consiste à approcher la fonction d'ajustement ou l'estimateur « optimal » afin de réduire la complexité des calculs. Les méthodes dites de « pseudo-vraisemblance » qui consistent à remplacer la fonction de vraisemblance par une autre, plus facile à maximiser, utilisent cette stratégie. Les méthodes variationnelles (développées notamment autour de M. Jordan : www.cs.berkeley.edu/~jordan/) constituent une piste particulièrement intéressante et efficace. Nous avons utilisé cette approche dans nos travaux sur le modèle de mélange pour les graphes aléatoires. Dans ce cas, du fait de la structure de graphe, l'étape E de l'algorithme EM mène à des calculs intractables que l'approximation variationnelle permet de surmonter. Nous parvenons ainsi à analyser des graphes de plusieurs milliers de sommets alors que le maximum de vraisemblance est inatteignable et que les méthodes MCMC ne peuvent pas dépasser 200 sommets. Si les méthodes variationnelles permettent d'utiliser des algorithmes très efficaces, elles ont deux défauts importants : on ne sait pas si la fonction qu'on optimise est éloignée de la fonction de vraisemblance d'une part et on connaît très mal les propriétés de convergence et de comportement (même asymptotique) des estimateurs. Nos travaux de simulation montrent que la qualité de l'approximation (et par conséquent celle des estimateurs) peut être très bonne dans certains cas, mais aussi assez mauvaise dans d'autres. Nous projetons d'avancer sur ces questions en essayant de préciser les conditions permettant d'espérer un bon comportement de ces méthodes, en prenant le cadre du modèle de graphe aléatoire comme exemple typique, mais avec des visées plus générales.

Une dernière piste consiste à explorer les liens entre inférence bayésienne et approximation variationnelle dans le cadre de la famille exponentielle. Cette combinaison doit permettre d'obtenir des intervalles de prédictions approchés, voire exacts des paramètres.

1.1.2 Initiation à la modélisation

« Cours » doctoraux

Nous comptons augmenter encore notre rôle de diffusion en modélisation statistique dans les laboratoires d'AgroParisTech et d'ailleurs. Nous avons écarté l'idée d'organiser une animation sous la forme d'un séminaire régulier sur le thème « statistique et biologie » qui viendrait s'ajouter à une liste déjà longue de séminaires parisiens.

Nous prévoyons de renforcer notre action de diffusion en s'adressant aux doctorants par des interventions accrues dans des écoles qui leurs sont destinées. Nous intervenons déjà dans le programme Erasmus « Math-bio » et dans un cours conjoint ABIES / Biologie Végétale (Orsay). Nous comptons également proposer des enseignements plus précisément liés à nos thèmes de recherche

- Analyse des données à haut débit en biologie
 - Modélisation hiérarchique en écologie et environnement
- dans le cadre de l'école doctorale ABIES ou en collaboration.

Monitorat : formation sous forme de stage

Comme décrit dans le bilan, nous accueillons déjà régulièrement des chercheurs en biologie (le plus souvent doctorant ou jeunes CR) qui viennent se former et conduire les analyses de leurs propres résultats dans notre unité. Pour l'instant, cette offre concerne principalement les biologistes moléculaires. Nous comptons étendre cette pratique (qui ne requiert pas de cadre statutaire particulier) à des doctorants issus de diverses domaines de la biologie (génétique, écologie, *etc.*).

L'apprentissage de la modélisation étant difficile sous forme magistrale, il nous semble plus efficace de le mener sous forme de stage court (1-3 semaines) au cours duquel les étudiants pourront s'approprier les modes de raisonnement et les méthodes statistiques dans le cadre de leur sujet de recherche.

1.2 Equipe « Statistique et Génome »

1.2.1 Vers des mesures à très haut débit

La technologie des puces à ADN, qui a maintenant une quinzaine d'années, se fonde sur l'hybridation complémentaire de deux brins d'ADN. A la naissance de cette technologie, les puces comportaient une centaine de sondes, puis grâce aux améliorations techniques, le nombre de sondes est passé à des dizaines de milliers, ce qui permet d'étudier simultanément tous les gènes d'un même organisme. La technologie continue encore de se perfectionner et on dispose maintenant de puces 'tiling array' dont les sondes couvrent l'intégralité du génome indépendamment de l'annotation structurale des gènes. Ce type de puce permettra de découvrir de nouvelles unités transcriptionnelles quand on hybridera des transcrits et de détecter des amplifications ou délétions de moins de 1kb quand on hybridera de l'ADN génomique (profil CGH).

Entre chaque type de puce, le nombre de données est multiplié par 100 et les questions biologiques s'enrichissent. D'un point de vue statistique, il est donc nécessaire de proposer à chaque fois des méthodes adaptées à la question et à la taille des jeux de données considérés. L'équipe s'est investie depuis plus de cinq ans dans l'analyse des données de puces et tout le travail réalisé lui permet d'être prête pour s'investir dans l'analyse des puces à très haute-densité, qui comporte de nouveaux enjeux méthodologiques car la quantité d'information contenue dans une expérience (presque 1 million de sondes) est telle qu'il est nécessaire de développer des méthodes efficaces et pertinentes pour réussir à extraire l'information et la synthétiser sous forme de résultats biologiquement interprétables.

Pour l'étude de données transcriptomes sur puces 'tiling-array', l'équipe a obtenu le financement de deux bourses de thèse pour travailler sur ce sujet. Les thèses débiteront en octobre 2008 et s'achèveront fin 2011. Nous avons choisi deux candidats ayant des compétences complémentaires pour couvrir l'analyse et la visualisation des données. Ces travaux seront menés en étroite collaboration avec l'URGV et aussi l'équipe de P. Bessières (MIG, INRA Jouy-en-Josas) impliqué dans le projet BaSysBio dont l'objectif est de découvrir la structure globale des réseaux de régulation du métabolisme bactérien.

Pour l'étude de profils CGH, les données générées avec des puces 'tiling-array' ont deux particularités : la première est que la taille des profils est beaucoup plus importante, passant de quelques centaines de points pour les CGH traditionnelles à quelques centaines de milliers de points avec ces nouvelles puces. La seconde est que les sondes utilisées pourront être chevauchantes, ce qui créera de la dépendance entre les points successifs d'un même profil. L'arrivée de cette technologie nécessitera donc à la fois le développement de nouveaux modèles ainsi que l'adaptation des algorithmes existants pour l'analyse des puces CGH traditionnelles. Sur ce dernier point, la principale difficulté concerne la segmentation de ces profils. En effet, les algorithmes de segmentation classiques ne pourront plus être utilisés. L'équipe a déjà commencé à travailler sur ce problème proposant un algorithme hybride (Gey & al, 2007) permettant de réduire au préalable le nombre de segmentations possibles en ne gardant que les plus pertinentes puis d'utiliser sur ces nouveaux candidats les algorithmes classiques. L'équipe continuera à travailler sur le développement de ce type d'algorithmes.

De plus, un ensemble de nouvelles technologies de séquençage à « très haut débit » a émergé depuis quelques années : 454, pyroséquençage, etc. Ces techniques permettent essentiellement d'échantillonner de courts fragments d'ADN (entre 25 et 15 nucléotides selon les méthodes) dans un échantillon. Le nombre de fragments est également variable mais atteint typiquement quelques millions de fragments par expérience.

Ces nouvelles technologies constituent ainsi une alternative à la plupart des applications des puces à ADN : transcriptome, « copy number », immuno-précipitation de la chromatine, etc. Ces techniques ouvrent également des horizons nouveaux avec les expériences métagénomiques qui visent à obtenir une vision globale de l'ensemble des génomes présents dans un même milieu (sol, intestin, ...).

Ces techniques sont réputées plus fiables que les puces pour 2 raisons principales :

- les mesures seraient entachées de moins de biais technologiques ;
- elles donnent accès directement à un nombre de copies (et non à un signal de fluorescence).

Si le premier argument reste à vérifier (certaines publications montrent des biais liés à la composition des séquences, au fait qu'elles soient répétées dans le génome, etc.), la seconde est importante d'un point de vue statistique puisqu'elle fait passer d'un signal continu (niveau de fluorescence ou de radioactivité) à un signal discret (comptage).

Nous nous proposons, dans un premier temps, de travailler à l'adaptation d'outils déjà disponibles pour les puces à ces nouvelles données. Ainsi, les méthodes d'analyse différentielle du niveau de transcription des gènes pourront être revisitées dans un cadre Poissonien (adapté aux données de comptages) au lieu du cadre Gaussien. Dans ce cas précis, nous pourrions profiter d'une analogie avec nos travaux sur les motifs dans les séquences d'ADN pour proposer des tests exacts ou asymptotiques, étudier leur puissance et en déduire des dispositifs expérimentaux pertinents.

De façon plus générale, nous travaillerons à l'extension de différents outils développés dans l'unité pour les puces (segmentation de signaux CGH, détection de gènes, analyse d'expériences de CHIP) aux techniques de

séquençage de cette nouvelle génération. Il est probable que certaines extensions ne demanderont que des adaptations directes, alors que d'autres demanderont des développements méthodologiques spécifiques. Du fait de la taille de données (10^6 fragments par expérience), les adaptations algorithmiques seront également nécessaires.

Nous commençons déjà à nous impliquer dans des projets concernant l'analyse de données issues de ces nouvelles technologies, notamment des expériences métagénomiques. Ces expériences visent, entre autres, à dénombrer, identifier, voire quantifier les espèces présentes dans un milieu. L'identification est relativement aisée si les espèces présentes sont déjà connues (*i.e.* séquencées). Un membre de l'UMR est porteur et plusieurs membres de l'UMR sont impliqués dans le projet ANR accepté pour 2009-2011 Computational Biology for Metagenomics Experiments (CBME), en collaboration avec MIG et MIA (INRA Jouy). La génomique microbienne à grande échelle est un domaine où des progrès scientifiques importants et rapides sont attendus, avec à la clé des applications dans les domaines de la chimie durable, la bioconversion et la santé. C'est un domaine où les possibilités ouvertes par les technologies de séquençage à haut débit sont particulièrement bien adaptées. Cependant, la technologie ne résout pas tous les problèmes, et il y a des obstacles méthodologiques. Au delà des problèmes de gestion de bioinformatique, la présence de quelques espèces très abondantes risque de cacher des espèces plus rares mais qui pourraient se révéler plus intéressantes. Cette question est à relier au nombre d'échantillons analysés : plus ils sont nombreux plus on a de chance de trouver des espèces (ou des gènes ou des fragments longs) pertinents...mais plus l'étude coûte cher. Il y a donc un problème de plan d'expérience à optimiser en fonction de ce que l'on sait de la répartition de l'abondance des espèces, mais aussi de la puissance des tests statistiques. De plus, l'analyse statistique elle-même pose des problèmes nouveaux: elle est rendue difficile par 3 éléments :

- (1) la multiplicité des tests ($\approx 10^8$) de comparaison entre conditions,
- (2) il y a beaucoup plus de variables analysées ($\approx 10^8$) que d'individus (≈ 100)
- (3) on a des données de comptages, c'est à dire des données discrètes, alors que les analyses classiques sont faites sur des données continues.

L'objectif du projet CBME est de fournir des outils bioinformatiques et statistiques adaptés aux expériences de métagénomique et de donner des éléments pour déterminer le nombre nécessaire d'échantillons pour atteindre avec une probabilité raisonnable les objectifs scientifiques visés par l'étude. C'est un projet général qui ne cible pas une expérience particulière de métagénomique. Cependant ce projet bénéficiera de l'implication de plusieurs de ses membres dans des projets de métagénomique dans différents écosystèmes dont le projet ANR accepté MetaSoil dans lequel un membre de l'UMR est partie prenante.

1.2.2 Apprentissage

Contexte

L'apprentissage artificiel a été dominé ces 20 dernières années par le problème de l'induction à partir de bases de données supposées identiquement et indépendamment distribuées et déjà fournies.

Depuis quelques années, ce paradigme est remis en cause, à la fois pour dépasser les limites du cadre théorique existant et en raison de l'apparition de nouvelles applications : analyse de flux de données, de données spatialisées, nouveaux systèmes « pervasifs » distribués, limités et à longue durée de vie, tâches de recommandation, ...

De nouvelles questions théoriques et pratiques apparaissent donc. Par exemple :

- Comment passer de critères inductifs fondés sur l'hypothèse de données identiquement et indépendamment distribuées à des critères inductifs prenant en compte une dépendance entre les données et l'évolution possible de la dépendance cible entre variables de description et variables de prédiction ?
- Comment adapter les algorithmes "batchs" _ prenant en compte directement toutes les données supposées disponibles_ pour en faire des algorithmes "en-ligne", donc avec oubli en partie des données vues ?
- Comment tenir compte des ressources computationnelles limitées des apprenants et, en particulier, comment obtenir des algorithmes "anytime" capables de fournir une réponse, éventuellement imparfaite, à tout moment ?
- Comment les agents apprenants peuvent-ils prendre en compte leurs propres limitations ?
- Comment transférer les résultats d'un apprentissage à une autre situation d'apprentissage ?

C'est dans le cadre de ces nouvelles questions théoriques et de ces nouvelles applications que se place notre activité scientifique et les projets envisagés. Les travaux récents ou en cours portent sur l'analyse dynamique des réseaux sociaux, le docking de protéines, l'analyse de données séquentielles par inférence d'automates et de grammaires, le transfert entre apprentissages, et l'apprentissage en-ligne. Les applications étudiées portent sur l'apprentissage en présence de dérive de concepts et en présence de "covariate shift" (contrat avec l'entreprise A2IA, co-encadrement de stagiaires chez EDF) et la mise au point de méthodes d'apprentissage pour l'analyse de flots de données et pour les systèmes pervasifs (action européenne KDubiq). Un projet est en cours de définition

pour l'analyse de données temporelles et spatiales relatives à la circulation de polluants dans l'atmosphère basse (projet ONU).

Apprentissage en ligne

La théorie statistique qui sert de cadre conceptuel à l'apprentissage présuppose l'existence d'une base de données d'apprentissage tirée indépendamment suivant une distribution fixe caractérisant également les données test sur lesquelles sera mesurée la performance de l'apprenant. Que se passe-t-il si l'apprenant n'a pas les ressources computationnelles lui permettant de traiter toutes les données d'apprentissage d'un seul coup ou si les données ne sont pas indépendantes ou si la distribution des données évolue en cours du temps ?

Une première interrogation concerne la possibilité d'adapter le cadre statistique classique à ces questions. Plusieurs directions sont étudiées dans la communauté scientifique. On distingue généralement le cas où seule la distribution $P(X)$ des exemples varie au cours du temps, et le cas où la dépendance du label par rapport aux covariables $P(Y|X)$ elle-même varie.

Concernant le premier cas ($P(X)$ variant au cours du temps), les études portent sur (i) l'inférence transductive (l'apprenant dispose d'une base d'apprentissage et des cas tests à l'avance) due à Vapnik, père de la théorie statistique de l'apprentissage, (ii) le covariate shift, dans lequel la distribution en apprentissage n'est pas la même que la distribution en test et (iii) le tracking dans lequel la distribution des exemples et des cas test évoluent au cours du temps. Dans les trois situations, l'enjeu est d'étudier comment modifier le critère inductif classique, fondé sur la minimisation du risque empirique, pour obtenir de meilleures performances.

Dans ce cadre, nos objectifs de recherche se focalisent en particulier sur l'étude du covariate shift, d'une part, et sur celle du tracking, d'autre part.

Le covariate shift ou dérive de la distribution des co-variables. Les méthodes d'apprentissage supervisé classiques supposent que la distribution des données est stationnaire. Or, soit pour des raisons liées à la constitution de l'échantillon d'apprentissage (ré-équilibrage des classes, apprentissage actif), soit parce que la distribution des exemples change au cours du temps, cette supposition peut se trouver erronée : c'est ce que l'on désigne par *covariate shift*. Il peut alors s'avérer judicieux de prendre en compte ce changement pour construire un classificateur qui devra être performant sur la nouvelle distribution des exemples X_{new} "probables".

Le covariate shift fait partie de nos sujets d'étude, tant en raison des projets applicatifs qui présentent cet aspect (projet A2IA, projet ANR soumis, projet ONU), que parce qu'il présente un cas particulier de modification possible du critère inductif pour tenir compte des évolutions temporelles des données et qu'il s'inscrit donc dans l'étude plus générale de l'apprentissage en-ligne. Le covariate shift représente une extension théorique assez directe des travaux déjà réalisés par les membres de l'équipe travaillant sur l'apprentissage statistique.

Le sujet étant très récent (cf. Workshop "*Learning when test and training inputs have different distributions*" à NIPS-06) et combinant des aspects de théorie statistique et informatique, l'UMR-MIA possède les compétences nécessaires pour le développement à la fois théorique et pratique de nouvelles méthodes d'apprentissage adaptées au covariate shift. En particulier, plusieurs pistes sont envisagées :

- L'étude de la minimisation du risque empirique après repondération des différentes observations dans le calcul de ce risque. L'ensemble des algorithmes basés sur la minimisation du risque empirique sont *a priori* facilement adaptables, pourtant peu d'articles se sont intéressés à cet aspect pratique. Notre équipe travaille déjà, bien que dans un cadre différent, à l'adaptation des algorithmes pour la minimisation d'un risque empirique pondéré. L'ensemble de nos travaux sur le sujet pourrait donc être étendu à la problématique du covariate shift.
- Il est toujours possible d'utiliser certains des critères pénalisés développés dans le cas classique, où le terme de pénalité ne dépend pas des données observées ou à venir (par exemple le critère d'Akaike ne tient compte que du nombre de paramètres nécessaires pour construire le classificateur). Mais dans le cas classique il a été observé que les critères adaptatifs (prenant en compte les données) donnent souvent de meilleurs résultats en pratique. Ainsi, les critères adaptatifs étudiés par notre équipe dans le cas classique (comme le critère swapping ou la validation croisée) pourraient être adaptés pour la prise en compte du covariate shift, et ce d'autant plus qu'ils semblent apparentés aux critères proposés pour l'inférence transductive dont on sait qu'elle est liée au covariate shift.

Le tracking. Le tracking est un concept ré-introduit récemment par Sutton et al. (2007) pour désigner des situations dans lesquelles les données d'apprentissage et les données test évoluent au cours du temps, même si la dépendance $P(X, Y)$ reste stationnaire. Il a été montré qu'avec des algorithmes d'apprentissage locaux (dans X) et simples, il était possible, sous ces conditions, d'obtenir de meilleurs résultats que des algorithmes classiques d'apprentissage. Cependant, le cadre théorique de ce type d'apprentissage reste à établir, et en particulier le lien entre mémoire du passé (traduite grâce à des paramètres utilisés par l'apprenant), évolution des données (vitesse de déplacement dans X) et performances (en généralisation sur des données locales dans X en fonction du

temps). Ici encore, le critère inductif est à réexaminer. Il devrait inclure des dimensions liées à la non stationnarité des données, à leur caractère dépendant et aux capacités computationnelles (essentiellement en espace mémoire) de l'apprenant. Par ailleurs, les applications présentant des opportunités de tracking sont nombreuses et pourraient bénéficier de critères inductifs adaptés et conduisant à des méthodes d'apprentissage plus performantes. On retrouve ici, sous une autre forme, le principe de l'induction transductive de Vapnik : il est inutile de chercher à prédire partout si l'on sait à peu près où se trouveront les exemples test. Les domaines d'application incluent la prédiction de mesures par des sondes dérivantes par exemple.

Le transfert entre apprentissages. Parce que l'apprentissage est de plus en plus considéré comme une activité continue et à longue vie ("Long-Life Learning" en anglais), la capacité à réutiliser des connaissances acquises lors d'un apprentissage pour un apprentissage ultérieur, éventuellement dans un autre domaine, fait l'objet d'un nombre d'études croissant. Ces études sont pour le moment essentiellement d'ordre heuristique, propre à chaque domaine et méthode. Afin de mettre en évidence des propriétés plus fondamentales sur la possibilité du transfert, nos travaux ont porté sur des situations génériques de tâches de recherche de chemin dans un graphe. Une grande partie des situations de résolution de problème peuvent en effet être vues comme des tâches de recherche d'un chemin. Les résultats obtenus [Fedon-et-al, 2008] semblent montrer que le transfert ne peut être intéressant que si des conditions très strictes sont vérifiées. Nous prévoyons de généraliser ces résultats afin d'en tirer des conclusions s'appliquant à de larges classes de situations.

Apprentissage de séquences. Une partie de l'activité des informaticiens de l'équipe concerne l'apprentissage de séquences, essentiellement sous deux aspects. D'une part, l'apprentissage de grammaires à partir de séquences a été étudié. Il a été montré [Pernot-et-al, 2009 ; Cornuéjols-et-al, 2008] que les algorithmes standards d'apprentissage de grammaires sont susceptibles d'être victime d'un phénomène de transition de phase déjà connu pour l'apprentissage de programmes logiques. Ce phénomène limite considérablement les capacités des méthodes actuelles. Son analyse théorique reste à compléter. Un livre est en préparation sur les phénomènes de transition de phase en apprentissage [Cornuéjols-et-al, Cambridge University Press, 2009]. D'autre part, nous nous intéressons aux fondements algorithmiques de l'induction à partir de séquences. A nouveau la question est d'identifier un critère inductif permettant de faire des prédictions à partir des observations passées. Mais ce critère ne repose plus directement sur des fondements statistiques, mais fait appel en particulier à des concepts de complexité algorithmique appliqués à des machines de Turing.

Intégration de données

L'évolution des technologies à haut débit permet maintenant de mesurer chez un même organisme le niveau d'expression des gènes, le nombre de copies d'ADN et le niveau de quantification de méthylation de l'ADN. Des méthodes existent pour analyser chaque type de données séparément, mais de plus en plus les biologistes souhaitent intégrer l'ensemble de ces données dans une analyse globale, afin de mieux comprendre les mécanismes impliqués. Cette nécessité de prendre en compte des données de nature différentes présente plusieurs difficultés pour le statisticien :

- proposer une modélisation satisfaisante des relations entre les différents types de données : par exemple, comme relier le niveau d'expression et le nombre de copies d'un même gène dans une cellule,
- proposer des algorithmes efficaces pour l'inférence des modèles proposés,
- proposer des procédures de test ou de sélection de modèles pour améliorer l'inférence et l'interprétation des résultats.

L'équipe a commencé à investir dans les méthodes d'intégration de données. En 2007 a commencé la thèse de Guillem Rigau, portant sur le développement de méthodes pour l'analyse simultanée des données de puces CGH et transcriptome pour la caractérisation des cancers du sein en vue d'identifier de nouvelles cibles thérapeutiques. L'étape de modélisation des données est actuellement en cours et se fait en collaboration avec Thierry Dubois (biologiste à l'Institut Curie, Département de Transfert, équipe de signalisation). Plusieurs modèles de complexité croissante sont actuellement envisagés, leur mise en œuvre algorithmique restant le facteur limitant de la démarche. C'est pourquoi nous travaillons à la mise au point de méthodes génériques, basées sur la programmation dynamique ou sur des algorithmes de classification (CAH), permettant de réaliser l'inférence sur de grands jeux de données en un temps raisonnable. L'objectif à terme est de proposer une famille d'algorithmes permettant facilement d'intégrer des données de 2 ou plusieurs natures, mais aussi de traiter les données des technologies à très haut débit produisant plusieurs dizaines de millions de mesures par expérience (v. partie 1.2.3). Par ailleurs, ce problème d'intégration de données hétérogènes portant sur un même phénomène est également au centre de recherches en apprentissage artificiel. La technique du co-apprentissage notamment permet de faire collaborer des « experts » (systèmes d'apprentissage artificiels) s'appuyant sur des données différentes pour tirer parti de données disponibles non étiquetées et pour renforcer leur certitude dans leurs prédictions. Le champ émergent de l'apprentissage semi supervisé est également pertinent dans cette perspective. Nous prévoyons donc d'investiguer ces techniques afin, à la fois, de les appliquer aux données biologiques disponibles, mais aussi de participer à leur analyse théorique qui est encore incomplète.

1.2.4 Analyse des réseaux d'interaction

La compréhension des réseaux biologiques (régulation, interactions protéiques, relations métaboliques) constitue un des enjeux majeurs de la biologie moléculaire actuelle. Plusieurs questions se posent :

- (1) modélisation déterministe de petits réseaux,
- (2) inférence de relations (moins précises qu'un modèle déterministe) entre entités nombreuses (les sommets du graphe) à partir de données observées sur ces entités au cours du temps,
- (3) identification de groupes de sommets ayant une même topologie de relations,
- (4) classification de noeuds ou d'arcs
- (5) identification de motifs topologiques fréquents dans un graphe
- (6) diffusion d'information sur un graphe à topologie fixée
- (7) évolution de la topologie d'un graphe au cours du temps.

C'est un sujet novateur et sur lequel l'UMR est bien placée. Nous avons le projet d'avancer sur nos points de compétences, c'est à dire les points (3), (4) et (5).

L'équipe a déjà obtenu des résultats intéressants sur les points 3 et 4 (Cf Bilan Réseaux) et compte développer de nouveaux modèles de graphes aléatoires et des méthodes performantes d'estimation des paramètres, en relation avec l'étude des méthodes approchées (Cf ci-dessus).

Le travail se fera en coordination avec le groupe de travail SSBnet pour le point (5). SSBnet regroupe des membres des UMR 518, INRA MIG et Université d'Evry, et a été présenté dans la partie Bilan. En effet, le groupe a déposé un projet ANR Blanc en Mathématiques, appelé NEMO, qui vient d'être accepté et dont le porteur est Stéphane Robin. Le contenu de NEMO est brièvement résumé ci-dessous :

Recent technical advances have allowed the collection of data which are structured like networks. Many scientific fields are concerned by this network revolution such as physics, sociology and biology. The study of interacting particles, social agents or proteins has led to the emergence of the notion of complex systems, and a major recent discovery is that most observed networks share similar structural properties, like the 'scale-free', the 'small-world' or the modularity properties. Consequently many mathematical strategies have been developed to uncover the global structures of complex networks, through their topology which describes the organization and evolutionary principles of such networks (Newman et al. 02).

Another strategy to investigate structure is to focus on local subunits or building blocks of networks, also called network motifs. This question has particularly arisen in molecular biology, as it has been showed that network motifs constitute functional units which combine to ensure cell regulatory processes (Shen-Orr et al. 02). More than a pure structural role, motifs are conserved among species, suggesting that some local topologies are preferred from an evolutionary point of view (Wutchy et al. 03, Chen et al. 06, Papp et al. 03).

Biological networks are classically represented as graphs; in protein-protein interaction networks for instance, nodes represent proteins and edges their interactions, whereas in regulation networks, nodes are genes or transcription factors and oriented edges indicate regulation activities. The definition of network motifs may depend on the nature of the network under study. For instance they can be defined by a connected sub-graph with a fixed topology and/or fixed node labels, but the definition can be more flexible to search for common structures in different networks. Despite a variety of definitions for motifs, their identification lies on the assessment of their exceptionality to determine if one particular substructure is favored regarding others. Consequently the identification of network motifs requires the detection of motifs which occur more frequently than expected in a random graph model to be specified (Shen-Orr et al. 02). This problem has given rise to a new dynamics in probability and statistics.

Current methodologies are in three steps. The first step consists in determining $N_{obs}(\mathbf{m})$, the number of occurrences of a given motif \mathbf{m} in the observed graph. It requires efficient computational strategies to obtain the count in a reasonable time. Current software use an estimation of this number, based on sampling in the real graph (Kashtan et al. 04). To determine if this count is exceptional or not, one needs to calculate the probability $P(N(\mathbf{m}) \geq N_{obs}(\mathbf{m}))$ of observing so many occurrences in a random graph (so-called p -value). The second step is then to choose a relevant random graph model that will correctly fit the biological network. The last step is the effective computation of the p -value which requires the knowledge of the count distribution; it is a critical point because this distribution is still not known, even under the simple model of Erdős-Rényi¹. Actual solutions are based on heavy simulation studies and/or a Gaussian approximation of the count distribution (Shen-Orr et al. 02, Milo et al. 02) which is theoretically valid only under very restrictive hypothesis on the graph and on the motif.

Objectives

The **NEMO (NETwork MOTifs)** project aims at addressing the three previous steps required to detect if a given motif has an exceptional frequency in a given network. Two types of motifs will be first considered, each one depending on the biological feature under study: motifs with a fixed topology (*topological motifs*) and motifs with no fixed topology but with fixed labeled nodes (*colored motifs*). Generalization to several other types of motifs will then be considered, such as mixed or degenerated motifs. We will:

- Provide efficient algorithms to count such motifs in real large networks.
- Characterize the motif count distribution in the Erdős-Rényi random graph model.
- Provide more realistic random graph models for biological networks.
- Generalize some results on the motif count distribution to these new random graph models.
- Implement our methods into a software to study motif exceptionality in networks.

Apply the developed methods to the analysis of the protein-protein interaction network of *Drosophila melanogaster*

1.2.5 Sélection de modèles

Les problèmes de sélection de modèles se rencontrent dans la plupart des méthodes que nous développons : modèles de mélanges, sélection de variables, détection de ruptures, etc.

¹ The Erdős-Rényi model assumes that edges are independent and identically distributed according to a Bernoulli distribution with a constant parameter.

Critères pénalisés.

Notre équipe a acquis une bonne compétence dans le domaine de la sélection de modèles, et notamment dans l'utilisation de critères de type « contrastes pénalisés ». Nous comptons poursuivre ces travaux aussi bien d'un point de vue applicatif que théorique.

Pour de nombreux problèmes, les critères pénalisés existent mais nécessitent la calibration de constantes spécifiques à chaque problème. Dans l'analyse des données de grandes dimensions, la compression du signal constitue une étape souvent nécessaire pour rendre l'information intelligible, voire pour obtenir des classificateurs efficaces. Le choix du taux de compression est un problème de sélection de modèles. Dans ce cas, l'heuristique classique dite « de pente » fournit un moyen assez facile à mettre en œuvre pour déterminer de façon automatique les constantes de pénalisation.

Notre équipe a déjà travaillé sur le problème de sélection posé par les modèles de détection de ruptures. Pour ce problème, il a été montré que le critère BIC pouvait être utilisé à condition de l'adapter convenablement. Nous comptons poursuivre dans cette direction en définissant les modèles à un niveau plus fin (*i.e.* conditionnellement aux positions des ruptures) que dans l'approche habituelle. Contrairement au BIC classique, cette approche aboutit à une pénalisation qui dépend de la loi *a priori* sur les modèles.

Ré-échantillonnage.

Le ré-échantillonnage désigne une famille d'algorithmes basés sur une heuristique visant à approcher la loi inconnue des observations par une loi obtenue à partir de nouveaux échantillons. Parmi ces algorithmes figure la validation croisée, largement utilisée dans de nombreuses communautés car elle fournit, en pratique, de très bons résultats. Son utilisation pose cependant des problèmes à la fois pratiques (temps de calcul) et théoriques (propriétés statistiques des estimateurs ainsi obtenus).

Notre équipe a contribué à faire progresser ces méthodes en proposant des formules closes (qui réduisent drastiquement le temps de calcul) et des résultats de consistance pour toutes une série d'estimateurs par projection utilisés en estimation de la densité ou en régression. Nous comptons poursuivre ces travaux notamment pour l'étude des régressogrammes, utilisés en détection de rupture. Dans le cadre de la régression, le ré-échantillonnage offre une alternative raisonnable aux critères pénalisés en terme de performances tout en demeurant valide sous des hypothèses moins restrictives : cadre hétéroscédastique et sans hypothèse forte sur la loi des résidus. Cependant si des résultats théoriques existent pour un nombre raisonnable de modèles, le cadre spécifique de la détection de ruptures rend ces derniers inapplicables. Un enjeu important consiste à mieux comprendre le fonctionnement de ces algorithmes dans ce cadre et à donner des garanties théoriques de validité de ces méthodes afin de s'assurer de la validité des résultats produits.

Plus généralement, les méthodes de ré-échantillonnage peuvent être vues comme des critères pénalisés avec une pénalité aléatoire. Outre les avantages ci-avant cités, les pénalités aléatoires semblent toutefois plus sensibles que leurs analogues déterministes au nombre de modèles mis en compétition : comparer un trop grand nombre de modèles « ressemblants » induit un phénomène de sur-ajustement, moins visible avec des pénalités du type de AIC. Un travail essentiel de compréhension des mécanismes reliant les pénalités aléatoires à la richesse de la collection de modèles est donc nécessaire.

1.3 Equipe MORSE

1.3.1 Optimisation de dispositif

(cf. Parent & al., 06 ; Parent & al., 08 ; Doukhan & al., 08)

Les modèles hiérarchiques développés en environnement (descriptions de pluies, de champs de vents ou de nuages de pollution, type ozone) visent à représenter les corrélations spatiales et temporelles entre sites de mesures. Nous avons initié une recherche pour jauger des difficultés méthodologiques et pratiques en matière de design optimal. En effet, utilisés en mode prédictif, ces modèles pourraient servir de base pour répondre à des questions d'ingénierie. Imaginons par exemple que l'on doive supprimer 50% de 100 stations pour des raisons budgétaires. Lesquelles doit-on supprimer pour minimiser la perte d'information selon l'objectif d'aménagement à considérer, par exemple :

- Une meilleure estimation statistique (maintenir une bonne précision globale des prévisions sur une ensemble de sites cibles (tels les aéroports pour les prévisions de brouillard) ou l'évaluation d'une quantité moyenne (par exemple la lame d'eau mensuelle reçue sur un bassin-versant),
- une caractéristique extrême (par exemple un pic de pollution, un minimum de débit dans les mois d'été permettant l'irrigation, ou la pluie maximale annuelle afin de dimensionner un système d'alerte aux crues.

D. Makowski, chercheur à l'INRA en agronomie doit venir travailler dans notre unité pendant un an pour travailler sur des « Méthodes Monte Carlo séquentielles pour améliorer les performances des modèles de culture dynamiques ».

1.3.2 Processus ponctuels

Les processus ponctuels et les processus ponctuels marqués constituent un formalisme adapté à la représentation de nombreuses données environnementales (position des arbres, des proies, clustering de proies, etc ...). Les méthodes d'estimation pour ce type de modèles sont bien connues dans les cas simples (processus poissons homogènes) mais restent assez imprécises dès qu'on souhaite modéliser des phénomènes plus compliqués (tendance, agrégation, ...). Nous chercherons à développer les méthodes d'estimations pour des processus ponctuels markoviens (de type Strauss, par exemple) ainsi qu'à compléter le test de caractère poissonien en calculant la loi asymptotique de la statistique de Ripley sous ces modèles afin de constituer une contre-hypothèse et d'étudier la puissance du test. Il est tentant d'utiliser ce type de modèle dans un cadre hiérarchique pour modéliser un processus d'intérêt (banques de proies par exemple) caché. Dans ce cadre, on manque cruellement de techniques d'estimation efficaces. Il est donc intéressant d'investir ce domaine de recherche notamment en utilisant des idées de conjugaison de processus Gamma Poisson.

1.3.3 Simulation conditionnelle et champs géostatistiques pour les extrêmes

Dans le but de proposer des alternatives au krigeage pour la prévision de processus spatiaux, des modèles de champs spatiaux doivent être développés afin de servir de cadre à des simulations conditionnelles. Les processus environnementaux qui nous intéressent ont la particularité de présenter des structures de dépendance, y compris pour les valeurs extrêmes et les modèles classiques qui sont en général Gaussiens ne conviennent pas.

On se basera sur le modèle tempête qui est un modèle max-stable, introduit par Smith et étudié par Schlather pour proposer des extensions. Les travaux sur la dépendance asymptotique déjà réalisés permettront de caractériser le type de dépendance associé aux données et de calibrer le modèle qui servira aux simulations. Les simulations conditionnelles (qui respectent les valeurs des observations) permettront de calculer des probabilités de dépassements de seuils, ponctuellement ou localement sur un support de taille variable.

1.3.4 Régression fonctionnelle spatialisée

Le modèle de régression fonctionnelle proposé dans un premier travail soumis à publication, pour établir un lien entre les courbes de climat et la mesure de diversité génétique du hêtre est un modèle linéaire basé sur la décomposition de Fourier des courbes de température et de précipitations, avec une prise en compte de la spatialisation par modélisation des résidus.

Ce modèle ne prend pas en compte la dépendance des variables d'entrée ni leur caractère aléatoire.

Nous comptons y remédier en adaptant par exemples les techniques PLS à ce type de modèle. Les résultats théoriques n'ont pas été établis pour des bases de Fourier, et cet aspect devra être étudié, probablement pour d'autres types de bases telles que les bases d'ondelettes, qui permettent une plus grande parcimonie pour les coefficients.

1.4 ANR soumises en 2008

Projets acceptés.

- NeMo (*Recherche de motifs dans les réseaux biologiques*) : ANR blanc math.
 - Partenaires : INRA MIG, Stat & Génome (Evry)
 - UMR 518 : S. Robin (porteur) + JJ. Daudin + M. Koskas, ETP = 47 mois sur 3 ans
 - Budget total de 120 k€ dont 72 k€ pour l'UMR.
- CBME (*Méthodes Bioinformatiques et Statistiques pour les expériences de Métagénomique*) : ANR génomique
 - Partenaires : INRA MIG, INRA MIA (Jouy)
 - UMR 518 : J.J. Daudin (porteur) + S. Robin + T. Maru-Huard + E. Lebarbier + J. Aubert, ETP = 36 mois sur 3 ans
 - Budget total 250k€ dont 120 k€ pour l'UMR
- MétaSol (*Description et exploitation de la communauté microbienne du sol par une approche metagénomique*) : ANR génomique
 - Ecole centrale de Lyon (T. Vogel) + 10 équipes partenaires
 - UMR 518 : S. Robin + E. Lebarbier, ETP = 11mois sur 3 ans
 - Budget total demandé de 2 400k€ dont 10 k€ pour l'unité

Projets rejetés.

- SMACC (*Statistical Methods for Array CGH and Cancer*): ANR Blanc math. Partenaires: CNRS / Curie.
- GENELOU (*Génétiq ue de système et syndrome métabolique chez le Rat*) : ANR génomique. Partenaire : INRA Bordeaux
- ROSA (*Bases génétiques du contrôle de l'activation de la motilité des spermatozoïdes en réponse à la salinité chez le tilapia *Sarotherodon melanotheron heudelotii**) : ANR génomique. Partenaire : IRD Montpellier.
- WISHUP : Wiki Intelligent Search for Unwanted edits Patrol : ANR Contint. Partenaires : Thales Research & Technology (coordinateur), AgroParisTech, CEA LIST, CRIHAN (Centre de Ressources Infomatiques de HAute-Normandie), PROXEM, Wikimedia France
- ExtremPat : ANR Blanc Math. Partenaire : Montpellier 2, CEMAGREF, Mines
- NANIMAT(Natural variation of nitrogen metabolism in *Arabidopsis thaliana*) : ANR Génomique. Partenaires : Unité Nutrition Azotée des Plantes, INRA Versailles.

1.5 Recrutements

Du fait du passage à l'INRA d'un professeur, le département MMIP devra recruter au printemps prochain un professeur de statistique pour AgroParisTech. De plus, d'ici à 5 ans, 3 enseignants de statistique du département MMIP partiront en retraite : un professeur fortement impliqué dans l'UMR 518 et deux maîtres de conférences n'ayant pas d'activité de recherche. Dans l'hypothèse (très vraisemblable du fait des charges d'enseignement du département) d'un maintien des effectifs, le département devra donc recruter 4 enseignants-chercheurs en statistique d'ici 5 ans. L'UMR verra ainsi son effectif augmenter de 3 enseignants-chercheurs.

Les profils de ces recrutements seront rédigés en fonction des impératifs d'enseignement et des orientations de recherche. Il est difficile d'indiquer d'ores et déjà quelles seront ces orientations. Elles seront évidemment liées aux axes de recherches indiqués dans ce projet, mais resteront aussi assez large pour permettre des recrutements plus opportunistes qui seraient l'occasion d'importer dans notre unité de nouvelles thématiques.

Les recrutements d'enseignants-chercheurs en statistique ne couvriront cependant pas tous les besoins de l'unité. Nous présentons ici deux besoins prioritaires.

1. Nos travaux sur les graphes, sur l'inférence de modèles complexes ou sur les données de très hautes dimensions posent tous des problèmes algorithmiques difficiles. La culture des statisticiens dans le domaine de l'algorithmique, des mathématiques discrètes ou de l'optimisation n'est pas suffisante. Un chercheur déjà expérimenté (CR1 / DR2) apporterait le renforcement méthodologique dont nous avons besoin. Nous n'espérons pas de recrutement sur un tel profil du côté d'AgroParisTech car ce besoin n'existe pas significativement du côté de l'enseignement.
2. Le développement et la diffusion d'outil font partie des missions première de notre unité. Ce travail passe généralement par le développement de logiciels efficaces. Les méthodes MCMC ou les calculs en grande dimension requièrent des compétences avancées en développement. Un ingénieur (d'étude ou de recherche) spécialisé dans le code numérique serait donc nécessaire pour compléter le travail des ingénieurs statisticien.

2 – Hygiène et sécurité

Identification des problèmes à résoudre au cours du prochain contrat et moyens envisagés pour y parvenir.

3 – Unités sensibles

Pour les unités sensibles (définies par le Fonctionnaire de Sécurité de Défense), seront pris en compte la capacité à protéger le patrimoine, la compréhension des enjeux, le respect des procédures, la sécurité des systèmes d'information, la protection de la propriété intellectuelle.