Demande d'allocation de Recherche 2012
Initiative Doctorale Interdisciplinaire de l'IDEX Paris-Saclay

Le projet fait-il l'objet d'un co-financement avec une entreprise ? Non.

1. Titre de la thèse

Extraction automatique d'information à partir d'article scientifique pour la reconstruction de réseaux de régulations biologiques impliquées dans le développement de la graine chez *Arabidopsis Thaliana*.

2. Directeurs de thèse

Nom : *Zweigenbaum* **Prénom :** *Pierre*

Adresse : *LIMSI, Université Paris-Sud, 91405 Orsay*

Téléphone : *01 69 85 80 04* Mail : *pz@limsi.fr*

Doctorat (année obtention): *1985* HDR (année obtention) : *1998*

Organisme de rattachement : *Directeur de recherche au CNRS*

Nombre de thèses en cours encadrées par le directeur de thèse (au moment de la demande) :

x 5 pour un taux d'encadrement total de 290 %, dont 1 (100 %) soutient en juin 2012 et 2 autres (60+20 %) soutiendront d'ici la fin 2012.

Nom : *Nédellec* **Prénom :** *Claire*

Adresse : *MIG, INRA, Domaine de Vilvert, 78352 Jouy-en-Josas*

Téléphone : *01 34 65 28 78* Mail : *claire.nedellec@jouy.inra.fr*

Doctorat (année obtention): *1994* HDR (année obtention) : *inscription en cours*

Organisme de rattachement : *Chercheuse à l'INRA*

Nombre de thèses en cours encadrées par le directeur de thèse (au moment de la demande) : 2

3. Nom de l'Unité - Laboratoire d'accueil

LIMSI : Information, Langue Ecrite et Signée. Traitement du Langage Parlé. Audio & Acoustique. Architectures et Modèles pour l'Interaction. Cognition, Perception et Usages. RV&A VENISE. Aérodynamique Instationnaire : turbulence et contrôle. Convection et Rotation : instabilités et turbulence Transferts Solide-Fluide

MIG : Bibliome, Biologie des Systèmes, Genevol.

Le travail reposera également sur une collaboration avec l'équipe de biologistes *Développement et qualité des graines* de l'Institut Jean-Pierre Bourgin (IJPB INRA, Versailles).

4. Les cinq dernières publications du directeur de thèse

1. L. Deléger, C. Grouin, et P. Zweigenbaum. Extracting medical information from narrative patient records: the case of medication-related information. *J Am Med Inform Assoc*, 17:555-558, 2010.
2. A. Ben Abacha et P. Zweigenbaum. Automatic extraction of semantic relations between medical entities: a rule based approach. *J Biomed Semantics*, 2(Suppl 5):S4, 2011.
3. C. Grouin et P. Zweigenbaum. Une approche à plusieurs étapes pour anonymiser des documents médicaux. RSTI-RIA, 25(4):525-549, *Numéro spécial Intelligence Artificielle et santé*. Hermès-Lavoisier, 2011.
4. A.-L. Minard, A.-L. Ligozat, A. Ben Abacha, D. Bernhard, B. Cartoni, L. Deléger, B. Grau, S. Rosset, P. Zweigenbaum, & C. Grouin. Hybrid methods for improving information access in clinical documents: Concept, assertion, and relation identification. *J Am Med Inform Assoc*, 18(5):588-593, 2011.
5. A. Ben Abacha et P. Zweigenbaum. Une étude comparative empirique sur la reconnaissance des entités médicales. In *Traitement Automatique des Langues*, 53(1), 2012, sous presse.

5. Résumé de la proposition

Le projet de thèse porte sur l'extraction automatique de connaissances sémantiques relationnelles à partir d'articles scientifiques portant sur le développement de la graine de la plante modèle *Arabidopsis thaliana*. Les connaissances extraites permettront la reconstruction des réseaux de régulations en incluant les niveaux génétique et moléculaire, les facteurs environnementaux et les phénotypes associés. La multiplicité des entités biologiques impliquées et la complexité de leurs relations nécessitent une approche, applicable à grande échelle, et capable de les distinguer à partir de textes en langue naturelle. L'approche retenue est celle de l'apprentissage automatique supervisé, appliqué à des exemples préalablement représentés et normalisés grâce à une analyse linguistique automatique, en premier lieu celle de l'approche de noyau à base de graphe et de sémantique distributionnelle qui nous a permis d'obtenir des résultats très prometteurs dans la prédiction de régulations géniques chez les bactéries. Nous voulons étudier particulièrement ici la représentation de l'information linguistique (variation terminologique, dépendances syntaxiques) la plus appropriée pour optimiser l'algorithme d'apprentissage et l'adapter aux spécificités de la rhétorique scientifique de la biologie moléculaire. Ce projet est basé sur le développement d'une collaboration faisant intervenir trois établissements (INRA, CNRS et Université Paris-Sud) et deux disciplines identifiées comme importantes pour l'Idex et la future Université Paris-Saclay : STIC et Biologie Végétale, portées par les LabEx DIGIWORLDS et SPS, et inscrite dans les pôles de compétitivité System@TIC, Cap Digital et le RTRA Digiteo. Ce projet se situe à l'interface et doit contribuer à renforcer leurs interactions.

6. Champs disciplinaires de la thèse et mots-clés

Champs disciplinaires : informatique, intelligence artificielle, apprentissage automatique, traitement automatique des langues, biologie des plantes, développement de la graine, biologie moléculaire.

Mots clefs : extraction d'information, reconnaissance d'entités nommées, reconnaissance de relations, analyse syntaxique, analyse terminologique, machine à vecteurs support, apprentissage faiblement supervisé, réseau de régulation génique, Arabidopsis, graine, développement.

7. Introduction et contexte permettant de cadrer le sujet

Bibliome et extraction d'information. La littérature scientifique constitue un gisement de connaissances scientifiques de grande valeur, mais largement inexploité parce qu'uniquement sous forme textuelle. La croissance très rapide du volume de publications à un niveau mondial rend impossible une veille scientifique systématique. Il est nécessaire de doter les chercheurs d'outils semi-automatiques pour sélectionner, extraire et formaliser ces connaissances, qui seront ensuite confrontées avec des connaissances de sources variées. C'est un des objectifs de l'extraction d'information (EI). Le domaine de la biologie offre un cadre particulièrement intéressant par les enjeux qu'il représente. L'EI en biologie a été popularisée depuis 2005 pour sa contribution à la construction de réseaux de régulations à partir d'interactions géniques (voir par exemple (Nedellec, *LLL* 2005 ; Bossy *et al.*, *BMC Bioinformatics* 2012)). Ces connaissances présentent un intérêt scientifique considérable car elles constituent l'une des étapes clés dans la construction d'hypothèses et de modèles mathématiques qui permettent au biologiste de mieux comprendre le fonctionnement des mécanismes étudiés. Une partie critique de cette connaissance biologique n'est pas décrite dans des banques de données, mais uniquement dans des articles scientifiques en langue naturelle sous une forme complexe. Les données de régulation en particulier sont nombreuses, dispersées dans la littérature et de nature hétérogène. Leur nombre augmente exponentiellement et elles sont donc de plus en plus difficiles à appréhender de façon exhaustive en vue de leur synthèse et de leur exploitation pour la compréhension des réseaux de régulation.

Réseaux de régulation. La collaboration entre MIG et l'IJPB a permis d'identifier un nouveau défi méthodologique pour l'EI dans la modélisation des réseaux de régulations chez la plante modèle *Arabidopsis*. Le modèle étudié est le développement des graines qui repose sur des mécanismes moléculaires et génétiques complexes (ex. coordination du développement de plusieurs tissus, d'origines génétiques différentes, en interaction avec l'environnement) (North *et al.*, *Plant J.* 2010). Un réseau de régulations très complexe permet cette coordination et notre compréhension de ce réseau dans sa globalité reste limitée (Santos-Mendoza *et al.*, *Plant J.* 2008). En plus de l'intérêt cognitif, le modèle choisi présente de nombreux intérêts finalisés. Les graines accumulent de grande quantités de composés de stockage (sucres lipides, protéines) ainsi que des métabolites secondaires, qui déterminent leurs utilisations pour l'alimentation humaine et animale ainsi que pour l'industrie (Baud *et al.*, *Plant Physiol Biochem* 2002). La graine est également le vecteur principal de l'amélioration génétique et de la production des grandes cultures annuelles. Une meilleure connaissance de la biologie et du développement des graines est donc un enjeu majeur pour l'agriculture et certaines

industries. Enfin, le développement des graines est important d'un point de vue écologique et pour l'évolution des espèces (résistance à la dessication, dissémination des espèces). Ceci justifie les efforts développés par un organisme comme l'INRA pour mieux comprendre les régulations qui contrôlent leur développement. Comme cela a déjà été montré pour plusieurs réseaux de régulation des grandes fonctions végétales, les résultats obtenus sur la plante modèle *Arabidopsis*, pourront assez simplement être transférés à l'étude de plantes cultivées.

Apprentissage supervisé pour l'extraction d'information à partir de textes. L'apprentissage automatique est l'approche dominante aujourd'hui en extraction d'information à partir de textes pour la biologie. La question de l'extraction de relations, formulée comme un problème d'apprentissage artificiel supervisé, consiste à générer automatiquement un classifieur à partir d'exemples de textes où les relations sont annotées manuellement. Le classifieur appliqué à de nouveaux textes peut alors prédire pour chaque paire d'arguments candidats préalablement identifiés, par exemple des gènes et des protéines, s'ils sont effectivement en relation (par exemple, *activation de l'expression*). La variabilité des formulations textuelles pour une même connaissance impose une étape d'analyse linguistique qui associe aux exemples d'apprentissage issus du texte, un ensemble d'informations linguistiques qui mettent en évidence des régularités, implicites dans le texte lui-même (Nédellec *et al.*, *Ontology Handbook* 2009). Les méthodes SVM (machines à vecteurs supports), ou dites à noyau, combinées avec des informations linguistiques (Tikk *et al.*, *PLoS Comput Biol* 2010), et les plus-proches-voisins (Culotta & Sorensen, *ACL* 2004) sont aujourd'hui les plus performantes. Parallèlement, les méthodes de régression logistique (Mintz, *ACL* 2009), ainsi que de réseaux de neurones (Barnickel, *PloS One* 2009) fournissent des résultats compétitifs dans des cas plus simples que le nôtre, comme l'interaction gène - protéine.

Verrous. L'application de l'EI à la modélisation de régulations chez *Arabidopsis* rencontre trois obstacles majeurs. (1) La connaissance à extraire est disséminée dans des articles complets, et non pas concentrée dans des phrases du résumé. Les entités impliquées doivent être distinguées finement (ex. gène, région promotrice, boites régulatrices) et de nouvelles entités doivent être prises en compte (facteurs environnementaux, hormonaux, phénotypes associés), ce qui rend la tâche de discrimination plus difficile. Les relations entre ces entités seront décrites à différents niveaux fonctionnels, génétiques, moléculaires et physiologiques. La complexité de la connaissance à extraire explique que les meilleurs résultats mesurés récemment sur des *benchmarks* de problèmes similaires plafonnent à 50% de rappel et précision (*BMC Bioinformatics*, 2012(13) *BioNLP special issue*). Pour traiter cette complexité tout en préservant les objectifs biologiques, nous proposons une approche originale basée sur un meilleur couplage de l'analyse linguistique et de l'apprentissage artificiel. La thèse étudiera une représentation des exemples pour l'algorithme d'apprentissage (graphe, arbre) qui soit le reflet de la sémantique profonde de la représentation linguistique et pas un appariement superficiel basé sur des similarités de forme (ex. arbre syntaxique / arbre de descripteurs).

(2) Le nombre élevé d'exemples annotés par les biologistes et nécessaire à l'apprentissage de connaissances complexes est un deuxième verrou critique. La meilleure exploitation de l'information linguistique que nous proposons le réduira en partie en normalisant les exemples. Nous voulons le réduire encore en développant une nouvelle approche combinant la supervision faible (*weak supervision*) (Riedel *et al.*, *EMNLP* 2010) et l'apprentissage actif (*active learning*).

(3) Enfin, l'extraction d'information de régulation chez *Arabidopsis* ne peut faire l'impasse du traitement d'articles complets qui soulève le problème bien identifié de l'analyse des coréférences. Il consiste à identifier automatiquement les différentes expressions (pronom, abréviation, *etc.*) qui dénotent une même entité dans un document, de manière à fusionner les connaissances reliées aux occurrences d'une même entité. Les résultats obtenus par les deux équipes STIC sur la résolution de coréférences anaphoriques (Zweigenbaum *et al., RFIA* 2012) seront adaptés par la thèse.

Compétence spécifique des équipes. Les équipes *Bibliome* et *ILES* possèdent la double compétence en traitement automatique des langues et en apprentissage artificiel nécessaire à l'étudiant. Les deux équipes collaborent sur plusieurs questions d'annotation sémantique automatique de texte dans le projet OSEO Quaero (Galibert *et al.*, *LREC* 2010). Les deux encadrants ont une compétence reconnue et acquise de longue date en extraction d'information, dans le domaine biomédical pour Pierre Zweigenbaum (Zweigenbaum *et al.*, *Brief Bioinform* 2007), et de la biologie moléculaire pour Claire Nédellec (Nedellec, *IEEE Intelligent Systems* 2002). Les deux équipes *Bibliome* et *Développement et qualité des graines* ont jeté les bases du problème d'extraction de régulation en formalisant ensemble le modèle d'extraction d'information et en constituant un premier corpus d'apprentissage représentatif. Leur compréhension réciproque de cette question interdisciplinaire est un gage de succès de la thèse. La collaboration étroite entre les deux équipes STIC et les biologistes de l'IJPB travaillant sur le développement de la graine permettra de considérer la question biologique dans toute sa diversité et sa complexité et sélectionner les mécanismes et données biologiques à analyser en priorité. Au niveau international, il existe peu de collaborations

étroites dans ces trois disciplines à la fois. La configuration exceptionnelle de notre collaboration permettra de traiter conjointement les questions de la représentation de l'information linguistique, des connaissances biologiques et de la performance de l'algorithme d'apprentissage, questions habituellement traitées séparément par les spécialistes de ces domaines. Nous sommes convaincus qu'elle permettra l'émergence de solutions originales.

8. **Méthodologie**

La recherche de l'algorithme d'apprentissage le plus performant pour l'extraction d'information est indissociable de la recherche sur la représentation des exemples la plus appropriée. La question complexe de l'appariement de la représentation linguistique et de la **représentation des données d'apprentissage** n'a pas été traitée de manière approfondie en raison de la nouveauté de la disponibilité d'analyses linguistiques de qualité et d'algorithmes d'apprentissage performants pour des représentations structurées (arbres, graphes). Les informations produites par l'analyse linguistique appartiennent à différents niveaux interdépendants de normalisation et d'abstraction : lemmes, étiquettes morpho-syntaxiques, dépendances syntaxiques, termes et variations, coréférences, classes sémantiques, concepts, rôles sémantiques. La question est alors, étant donné la richesse de l'information linguistique, de proposer une représentation des exemples et une technique d'apprentissage aptes à en tirer parti, tout en conservant des propriétés calculatoires raisonnables.

Deux types de connaissances linguistiques nous paraissent particulièrement critiques.

L'apport des **dépendances syntaxiques** est maintenant reconnu comme marqueur de rôles sémantiques, et donc des relations cibles. Par exemple, la protéine, sujet d'un verbe d'interaction dont l'objet direct est un gène, est probablement l'agent de l'interaction dont le gène objet est la cible. Diverses représentations des informations syntaxiques sont étudiées : séquences (Culotta *et al.*, *ACL* 2006 ; Bunescu & Mooney, *NIPS* 2006), *shallow parsing* (Pustejovsky, *PSB* 2002 ; Zelenko, *JMLR* 2003), arbres (Zhang, *et al* ; ACL 2006 ; Liu *et al*, *NAACL* 2007), graphes (Culotta & Sorensen, *ACL* 2004 ; Fundel, *Bioinformatics* 2006) et dans l'équipe Bibliome, chemins de dépendances entre les arguments de la relation (Manine *et al.*, *Int J Med Inform* 2009), mais les représentations proposées prennent peu en compte la variabilité des analyses syntaxiques pour exprimer une même connaissance, par exemple, l'ordre variable des arguments ou les ambiguïtés d'attachement syntaxique. Nous proposons d'aborder cette question par une étape de normalisation linguistique préalable à l'apprentissage, basée sur l'exploitation de la **sous-catégorisation des prédicats** nominaux et verbaux.

L'utilisation par l'apprentissage pour l'EI de **connaissances sémantiques induites automatiquement à partir de corpus** est encore débutante. Les résultats obtenus par la combinaison de dépendances syntaxiques avec la sémantique distributionnelle (utilisation de classes sémantiques de mots apprises automatiquement par *clustering* à partir de corpus) sont prometteurs (Sun *et al.*, *ACL* 2011). Notre méthode SPSK basée sur un alignement global qui repose sur la sémantique distributionnelle, pour un *dependency path kernel* obtient des résultats parmi les meilleurs évalués sur les données LLL d'interactions géniques (Zargayouna *et al.*, *rapport Quaero* 2011,). Nous proposons d'étendre SPSK en utilisant la sémantique distributionnelle pour compléter l'apport des cadres de sous-catégorisation en induisant des proximités sémantiques qui rendent compte, non seulement de la proximité sémantique des termes du domaine, mais aussi de la variabilité terminologique et de la sémantique des dépendances syntaxiques et des prépositions.

Outre l'utilisation d'informations linguistiques pour maîtriser la complexité de la représentation, la rareté des exemples d'apprentissage sera compensée par une approche **faiblement supervisée** (Hearst, *COLING* 1992 ; Craven & Kumlien, *ISMB* 1999 ; Warnier *et al.*, *BioNLP* 2011). Notre proposition consiste à exploiter les bases de données disponibles sur les régulations chez *Arabidopsis* pour préannoter automatiquement les exemples considérant que quand les arguments de la relation sont co-cités dans un texte d'entraînement, alors la relation est *probablement* dénotée par le texte. Notre nouvelle approche de la supervision faible modifiera incrémentalement et en temps réel les règles de projection de la base de données sur le texte au fur et à mesure du *feedback* des annotateurs de l'équipe de l'IJPB, de manière à traiter le problème critique des faux positifs. Nous utiliserons l'éditeur d'annotation AlvisAE (Bossy *et al.*, *LAW ACL* 2012) développé par l'équipe Bibliome. Deux types de raffinement seront explorés. En cas d'ajout de nouveaux exemples positifs, des règles de variations terminologiques associeront les synonymes et paraphrases éventuelles. En cas d'invalidation de l'annotation automatique, un classifieur de sac de mots filtrera les phrases qui n'expriment pas la relation recherchée. Notre méthode appliquée à ce problème chez les bactéries obtient déjà une F-mesure de 80 % (Warnier *et al.*, *BioNLP* 2011).

9. Résultats attendus et perspectives

Le doctorant évaluera et publiera sa méthode en participant aux challenges d'extraction d'information en biologie, BioNLP et BioCreative. Les supports de publications seront les conférences en apprentissage (CAP, KDD, ECML/PKDD, ICML) et en traitement automatique des langues (ACL, Coling, EMNLP, LREC), les workshops et journaux sur les applications de l'informatique à la biologie (BioCreative, BioNLP, BMC Bioinformatics, BMC System Biology) et les conférences et journaux en intelligence artificielle (RFIA, ECAI, IJCAI, International Journal of Human-Computer Studies). Les résultats originaux obtenus en biologie des plantes pourront être publiés dans les journaux de ces domaines : Plant Journal, Plant Cell., ou dans des revues plus généralistes, PNAS, Current Biology.

10. Calendrier prévu pour les 3 années

Dans une 1re étape, l'étudiant s'appropriera les méthodes basées sur des noyaux. Il évaluera les apports des différentes représentations linguistiques comparativement aux autres méthodes de l'état de l'art. En parallèle, il affinera la définition de la tâche avec les biologistes de l'IJPB. L'approche *faiblement supervisée* sera développée dans un premier temps pour limiter le nombre d'exemples à annoter par l'IJBP. Une fois le corpus d'apprentissage obtenu, la 2e étape de la thèse proposera une représentation des exemples et un algorithme d'apprentissage capable d'exploiter l'information de la sémantique distributionnelle et des cadres de sous-catégorisation. Elle sera évaluée sur les phrases du corpus d'*Arabidopsis* et sur les *benchmarks* adaptés, qui ne nécessitent pas de traitement de coréférence, en particulier BioNLP'11 et BioNLP'13. Les résultats donneront lieu à publication. Le traitement de la coréférence sera intégré dans une 3e étape. La méthode sera alors appliquée à grande échelle à la prédiction biologique pour *Arabidopsis*. Les résultats seront publiés en fonction de leur pertinence dans les journaux cités. La méthode pourra ensuite être généralisée à d'autres espèces, le blé avec le GDEC (INRA) et les bactéries avec MICALIS (INRA), grâce aux collaborations de Bibliome. Ce calendrier sera synchronisé avec le plan de formation de l'école doctorale.

11. Conditions de réalisation de la thèse

L'étudiant répartira son temps de manière équilibrée entre les équipes Bibliome de MIG et Langue Ecrite et Signée (ILES) du LIMSI. Dans l'équipe ILES, l'étudiant bénéficiera de l'expérience de l'équipe en traitement automatique des langues et plus spécifiquement en extraction d'information biomédicale, avec la reconnaissance d'entités et de relations, la détection de négations et de modalités, la détection d'anaphores, et plus largement le traitement de la paraphrase. L'expérience de l'équipe Bibliome en apprentissage artificiel, analyse terminologique et annotation manuelle de corpus sera mise à profit ici. Les recherches plus particulièrement coordonnées avec cette thèse sont en sémantique distributionnelle avec le LIG (UJF Grenoble) et en acquisition de ressources syntaxico-sémantiques avec le Lattice (ENS). Concernant les moyens techniques et matériels, l'étudiant en thèse bénéficiera de la plateforme d'apprentissage et d'annotation sémantique *Alvis* de l'équipe Bibliome pour les traitements linguistiques nécessaires à la production des exemples d'apprentissage. La méthode développée par l'étudiant sera implémentée sous la forme d'un module intégré à *Alvis* permettant son évaluation et son exploitation. L'étudiant pourra expérimenter ses algorithmes sur la plateforme bioinformatique Migale (IbiSA) de l'unité MIG (500 processeurs).

12. Bibliographie annexe (références des équipes Bibliome et de l'IJPB)

Baud S., Boutin J.P., Miquel M., Lepiniec L. and Rochat C (2002) An integrated overview of seed development in Arabidopsis thaliana ecotype WS. *Plant Physiol Biochem* 40, 151160.

R. Bossy, J. Jourde, A.-P. Manine, P. Veber, E. Alphonse, M. van de Guchte, P. Bessières, C. Nédellec. (2012) BioNLP Shared Task - The Bacteria Track. *BMC Bioinformatics*.

A.-P. Manine, E. Alphonse, P. Bessières (2009). Learning ontological rules to extract multiple relations of genic interactions from text. *Int. J. Med. Inform.* 78(12):31-38.

C. Nédellec, A. Nazarenko et R. Bossy (2009). Information Extraction. *Handbook on Ontology*. S. Staab, R. Studer (eds.), Springer Verlag, p. 662-687.

North H, Baud S, Debeaujon I, Dubos C, Dubreucq B, Grappin P, Jullien M, Lepiniec L, Marion-Poll A, Miquel M, Rajjou L, Routaboul JM, Caboche M. (2010) Arabidopsis seed secrets unravelled after a decade of genetic and omics-driven research. *Plant J.* Mar;61(6):971-81.

J.-P., Pois D., Tannier X., Deléger L., Laurent D., « Named and specific entity detection in varied data: The Quæro Named Entity baseline evaluation », *The seventh international conference on Language Resources and Evaluation (LREC-2010),* European Language Resources Association (ELRA) publisher, Malte, 19-21 mai 2010.

Z. Ratkovic, W. Golik, P. Warnier. (2012) BioNLP 2011 Task Bacteria Biotope - The Alvis System. *BMC Bioinformatics*.

Santos-Mendoza, M., Dubreucq, B., Baud, S., Parcy, F., Caboche, M., and Lepiniec, L. (2008). Deciphering gene regulatory networks that control seed development and maturation in Arabidopsis. *Plant J* 54, 608-620.

13. Avis du responsable de l'unité

MIG (INRA) Jean-François Gibrat (pour S. Schbath)

La thèse proposée s'appuiera sur les compétences d'équipes multidisciplinaires (les STIC dans leur composante apprentissage automatique et traitement automatique de la langue et la Biologie végétale) de différents établissements du plateau de Saclay : LIMSI CNRS et Université Paris Sud, MIG INRA Jouy et Institut J-P Bourgin INRA-Versailles. Cette thèse bénéficiera de collaborations déjà existantes (LIMSI-MIG dans le cadre du projet Quaero) et débutantes (MIG-IJPB sur les aspects développement des graines). Elle s'inscrit pleinement dans la problématique de recherche de l'unité qui consiste à développer des méthodes permettant d'acquérir des connaissances à partir des données biologiques et, plus particulièrement, pour l'équipe Bibliome, à extraire des informations biologiques d'intérêt (par exemple des régulations) à partir de la littérature scientifique. Les travaux de thèse reposeront sur des approches qui ont été testées avec succès dans le cadre microbien et qui seront étendues et généralisées au cas des plantes. En effet, dans ce dernier cas, les problèmes à résoudre sont beaucoup plus complexes et nécessitent la mise en oeuvre de nouvelles approches. Les équipes impliquées sont reconnues dans leurs domaines respectifs. En outre, elles disposent d'une excellente candidate pour la thèse qui a déjà commencé à travailler sur cette thématique dans le cadre de CDD. J'émets un avis très favorable pour ce sujet de thèse.

LIMSI (CNRS et Université Paris Sud) Anne Vilnat (pour P. Le Quéré), directeur adjoint

Ce projet de thèse réunit un laboratoire de l'INRA et le LIMSI-CNRS sur une thématique pluri-disciplinaire. Elle permettra de donner corps à une collaboration initiée dans le cadre du projet franco-allemand Quaero et de la prolonger au-delà du projet, qui se termine en 2013. Sur le plan scientifique et thématique, cette thèse s'inscrit dans l'activité du thème « Extraction et recherche d'information » du groupe ILES, plus particulièrement sur l'extraction de relations. Une thèse se termine cette saison sur cette problématique, et une nouvelle thèse va permettre d'explorer de nouvelles pistes et d'aller plus loin, en particulier du fait que le profil de doctorant souhaité (pour lequel une candidate a été trouvée) a des compétences spécifiques en apprentissage automatique. Le sujet a comme domaine d'application les sciences de la vie, domaine qui fait l'objet d'un intérêt particulier dans le groupe ILES et qui a été sélectionné par l'université comme BQR en 2011, sur lequel s'est fait un recrutement MCF en 2012. Cette interaction STIC-Sciences de la vie figure également dans le Labex DIGIWORLDS dans lequel le LIMSI est impliqué. Elle s'inscrit donc dans une dynamique forte de développement dans le laboratoire. Pour toutes ces raisons, le LIMSI apporte son soutien à ce projet qui entre dans des thématiques en plein développement au sein du groupe ILES.

14. Avis motivé du responsable de l'Etablissement :

INRA Jouy-en-Josas Muriel Mambrini, présidente du centre INRA de Jouy-en-Josas

Ce projet de thèse combine trois originalités majeures pour dépasser un verrou de taille dans le champs de l'extraction de connaissances pour la compréhension de la régulation des réseaux biologiques. Les approches développées à ce jour montrent leurs limites dès lors que la connaissance à extraire est plus complexe. La première originalité est de proposer des combinaisons inédites de i) méthodes d'analyse linguistique pour discriminer l'étendue des composantes des réseaux de régulation et ii) d'exploitation de l'information linguistique pour extraire à partir de sources étendues et réussir à fusionner. Il sera alors possible d'explorer de manière étendu le gisement de connaissances des publications. Ceci est rendu possible par la seconde originalité qui est une association relativement inédite de compétences en apprentissage automatique, en traitement de la langue et en biologie végétale. La troisième originalité, qui garantit le succès de l'entreprise, est l'équilibre de mise en jeu de ces compétences et leurs convergences temporelles permises par la méthodologie proposée. Tout en proposant un projet de formation par la recherche en extraction d'information et traitement automatique des langues, des allers retours subtils et séquentiels seront effectués dans le domaine de la biologie pour donner et accroître le sens et l'efficacité des méthodes que le doctorant produira. Ainsi développera-t-il sa propre problématique scientifique, environné par un socle affirmé de compétences d'experts déjà bien engagés dans des collaborations. Ceci est l'assurance d'une thèse innovante, qui aura des impacts dans chacun des domaines, qu'elle contribuera à rapprocher : les STIC et la biologie. Le doctorant est assuré de trouver les compétences, réseaux de relations et dispositifs dont il aura besoin. Ce projet est un moteur alimentant l'enjeu collectif du centre, qui est d'épauler le développement d'une biologie plus prédictive pour contribuer aux objectifs de l'INRA 2020, et répondre aux ambitions portées par l'Université Paris-Saclay. J'encourage une telle collaboration entre le CNRS, Paris Sud et les deux centres franciliens de l'INRA. Pour toutes ces raisons, j'émets un avis très favorable pour ce projet de thèse.