Extraction d'information pour la sélection du blé par marqueur génétique

Dialekti Valsamou, Robert Bossy, Marion Ranoux, Wiktoria Golik, Pierre Sourdille, Claire Nédellec

Extraction d'information pour la sélection du blé par marqueur génétique

Dialekti Valsamou[1], Robert Bossy[1], Marion Ranoux[2] Wiktoria Golik[1], Pierre Sourdille[2], Claire Nédellec[1]

[1] INRA, unité UR1077 MIG (Mathématique, Informatique et Génome),
Domaine de Vilvert, 78 352 Jouy-en-Josas

[1] INRA, unité UR1095 GDEC (Génétique, Diversité, Ecophysiologie des Céréales),
Site de Crouël, 5 chemin de Beaulieu, 63 039 Clermont-Ferrand cedex
{prénom.nom}@jouy.inra.fr

Résumé : La sélection des variétés de blé d'intérêt agronomique est un enjeu millénaire. Les évolutions récentes de la biologie permettent d'accélérer considérablement le processus de sélection par l'identification de marqueurs génétiques liés à des gènes impliqués dans le contrôle d'un caractère d'intérêt comme le nombre de grains ou la résistance aux maladies. Nous décrivons ici la méthode mise en œuvre pour extraire de façon automatique et à grande échelle, cette information de la littérature scientifique. Elle s'appuie sur la définition d'un modèle de connaissance adapté à sa représentation et à son extraction, ainsi que sur le thésaurus *WheatPhenotypes* des caractères et phénotypes construit dans ce but. La première étape de l'extraction automatique porte sur la reconnaissance des entités par deux méthodes complémentaires, l'une pour les termes figés basée sur des dictionnaires et patrons et l'autre pour les termes plus variables basée sur une analyse terminologique et sémantique. Elles sont évaluées sur des ensembles d'exemples annotés manuellement, de référence. Les premiers résultats d'extraction d'entités obtenus sont très prometteurs. Ils sont exploités dans le moteur de recherche sémantique *AlvisIRWheatMarker* spécialisé dans la recherche d'information de marqueurs génétiques du blé.

Mots-clés : *extraction d'information, analyse linguistique, annotation de corpus, construction d'ontologie, biologie, génétique*.

1 Introduction

L'extraction d'information à partir de textes a connu un développement très important ces dernières années dans ses applications à la biologie. Elles portent essentiellement sur des questions de biologie moléculaire telles que les interactions protéiques ou géniques et sont popularisées par les compétitions BioCreative [Hirschman et al., 2005], BioNLP [Kim et al., 2009] ou LLL [Nédellec, 2005]. Les besoins d'extraction d'information en biologie sont pourtant très divers. L'extraction des traits et phénotypes rencontrent un intérêt grandissant en raison à la fois de l'importance de cette connaissance [Collier et al., 2013], mais aussi parce que leur extraction pose de nouvelles questions méthodologiques dues à la variabilité terminologiques, à la diversité des types de connaissances impliquées à différents niveaux biologiques. La difficulté de la tâche d'extraction augmente avec le nombre de relations et de types d'entités. C'est le cas de l'extraction d'information pour la sélection du blé assistée par marqueur génétique dont nous décrivons ici la formalisation du problème, la mise en œuvre et les résultats préliminaires d'extraction.

2 Sélection du blé assistée par marqueur génétique

L'amélioration des espèces d'animaux et de plantes d'intérêt dans un futur proche est devenue un enjeu international en raison des forts besoins pour l'alimentation d'une population mondiale croissante. Les nouvelles contraintes comme la réduction des intrants (eau, fertilisants et pesticides) et des surfaces cultivées requiert le développement des nouveaux schémas de sélection plus courts et plus efficaces. Les marqueurs moléculaires fournissent un outil précieux pour une sélection indirecte des caractères, mais leur utilisation impose d'avoir au préalable identifié des liaisons entre ces marqueurs et des gènes impliqués dans l'expression des caractères d'intérêt agronomique. Les progrès des outils génomiques

contribuent à améliorer la connaissance sur la liaison entre marqueurs moléculaires et gènes d'intérêt agronomique. Cette information doit être intégrée dans des programmes de sélection avec le but de passer de la sélection génétique à la sélection génomique. Une grande variété et un grand nombre de marqueurs moléculaires ont été développés ces dix dernières années chez le blé tendre (pour revue voir [Paux et al., 2009]. Pourtant l'information connue la plus utile est publiée dans des articles et doit être extraite de dizaines de milliers d'articles dont seuls quelques uns sont pertinents. Dans chaque article pertinent, seule une petite partie traite réellement du sujet en indiquant le nom du marqueur le plus proche du gène d'intérêt, le gène lui même et le protocole qui est utilisé pour révéler le signal moléculaire qui peut être utilisé pour la sélection assistée par marqueurs.

L'objectif du travail rapporté ici est d'extraire à partir de textes non structurés en langue naturelle, les relations entre des caractéristiques simples comme les marqueurs moléculaires, les gènes, les traits, les phénotypes et les variétés. Les traits sont les caractères observables, comme la résistance à une maladie. Les phénotypes sont les valeurs de ces traits, comme la résistance ou la sensibilité dans le cas de la maladie. Le phénotype est contrôlé par des allèles de gènes (le génotype de l'individu). Les allèles sont différentes versions d'un même gène. A un allèle est souvent associé un marqueur moléculaire. Ce marqueur permet de discriminer les différents allèles d'un gène grâce au polymorphisme présent sur la séquence d'ADN.

Les marqueurs génétiques sont utilisés pour sélectionner les variétés qui présentent un phénotype d'intérêt agronomique. Les données de liaison entre marqueurs et gènes d'intérêt que nous étudions appartiennent à quatre grandes thématiques : (1) les stresses biotiques, c'est-à-dire les gènes de résistance aux maladies fongiques et virales (rouilles, oïdium, fusariose, septoriose…) et les gènes de résistance aux ravageurs (pucerons, cécidomie, nématodes, mouche de Hess…) ; (2) les stresses abiotiques (sécheresse, salinité, verse…) ; (3) le développement de la plante (photopériode, vernalisation, floraison…) ; (4) la qualité boulangère (dureté du grain, protéines de réserve, taux de protéines…).

Les principales difficultés pour l'extraction automatique sont les nombreuses façons d'exprimer une même conclusion (Exemple 1), et la densité élevée de relations impliquant différentes entités mentionnées (Exemple 2).

Exemple 1. La propriété de résistance à la rouille de la feuille (*leaf rust*) chez le blé par le gène *Lr34* est aussi bien décrite par

the gene Lr34 confers resistance to leaf rust que par

lines missing Lr34 allele are susceptible.

Exemple 2. [PMID 20002313]

only two alleles, photoperiod insensitive (Ppd-D1a and Ppd-B1a) and photoperiod sensitive (Ppd-D1b and Ppd-B1b), respectively, at each locus were known previously

Dans l'exemple 2, les quatre entités de type allèle (*Ppd-D1a*, *Ppd-D1b*, *Ppd-B1a*, *Ppd-B1b*) et les deux entités de type phénotype (*photoperiod insensitive et photoperiod insensitive)* sont les arguments de quatre instances de la relation *allele_expresses_phenotype*.

Les progrès récents de l'extraction d'information en biologie, mesurés dans les compétitions comme BioNLP Shared Task permettent d'envisager l'extraction systématique et à grande échelle avec une qualité qui satisfasse les besoins des sélectionneurs.

3 Corpus annoté et ontologie

La collection de documents utilisée pour l'extraction est composée des textes complets de 3 170 articles scientifiques, qui traitent des liens entre marqueurs moléculaires et les gènes d'intérêt, sélectionnés dans la production scientifique mondiale. Les articles ont été sélectionnés en interrogeant *Web of Science* (WoS) avec les mots-clefs *wheat*, *marker* et *gene* pour identifier les références et en interrogeant les sites des éditeurs pour les textes complets.

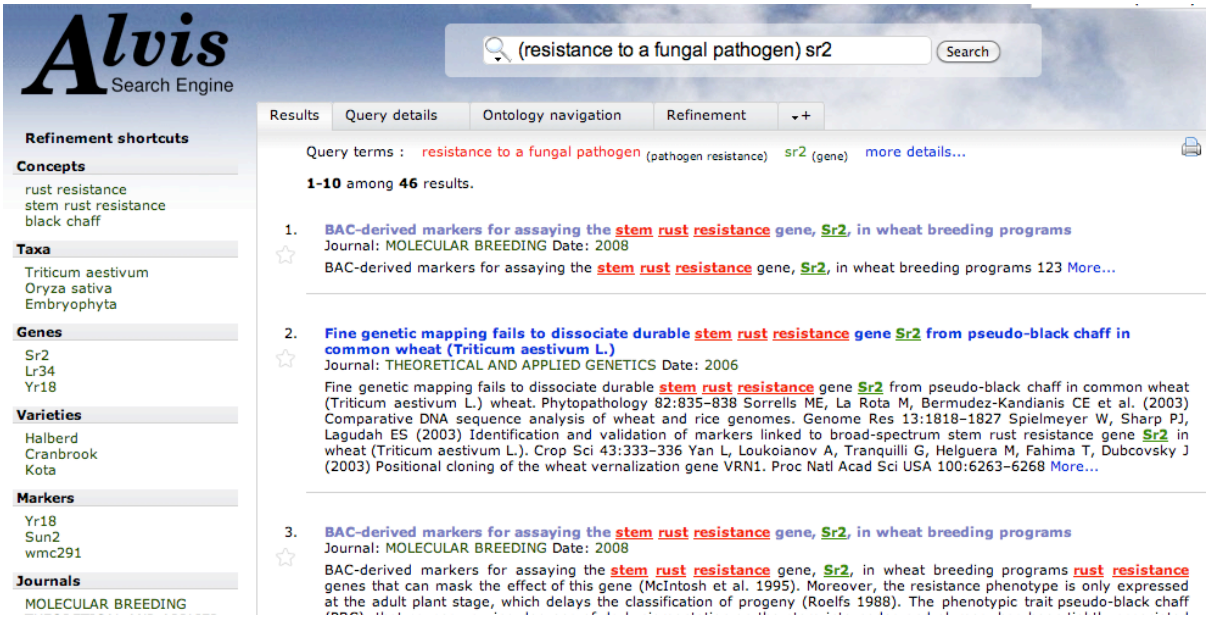


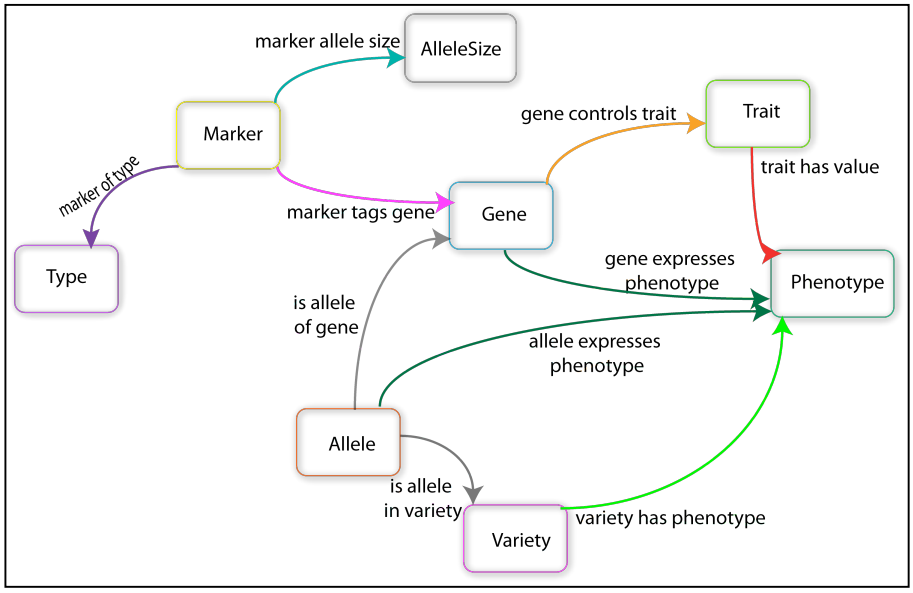
Figure 1. Interface du moteur de recherche sémantique AlvisIRWheatMarker

Avec les sélectionneurs impliqués dans le projet, nous avons construit un modèle original de connaissance pour représenter l'information pertinente des textes dans toute leur diversité, les marqueurs, les gènes, les variétés et les traits et phénotypes d'intérêt agronomique. Ce modèle est également utilisé par le moteur de recherche sémantique *AlvisIRWheatMarkers* qui indexe la collection documentaire. La figure 1 montre une copie d'écran de l'interrogation du moteur sur l'implication du gène *sr2* dans la résistance à une maladie fongique (*resistance to a fungal pathogen*). Les documents « réponses » de l'exemple mentionnent en particulier *stem rust resistance* qui est une spécialisation du concept *résistance à une maladie fongique*. Le moteur est accessible publiquement[1]. Une fois l'information extraite automatiquement et évaluée, elle sera intégrée dans une base de données accessible à distance et interconnectée avec les autres de données génomiques pertinentes comme la carte physique et ses 4000 marqueurs moléculaires et les séquences connues des chromosomes [Raats et al., 2013].

Le modèle du domaine comprend 14 relations n-aires (10 relations binaires et 4 relations ternaires) et 8 entités (Figure 2). Il a été nécessaire de définir des relations binaires partiellement redondantes avec des relations ternaires pour traiter les cas où l'information du texte est incomplète. Par exemple la relation *marker_tags_gene* subsume la relation *marker_tags_gene_in_variety*, mais elle est nécessaire quand l'information de variété de blé concernée est absente.

[1] http://bibliome.jouy.inra.fr/test/alvisir/FSOV/

a. Relations binaires



b. Relations ternaires

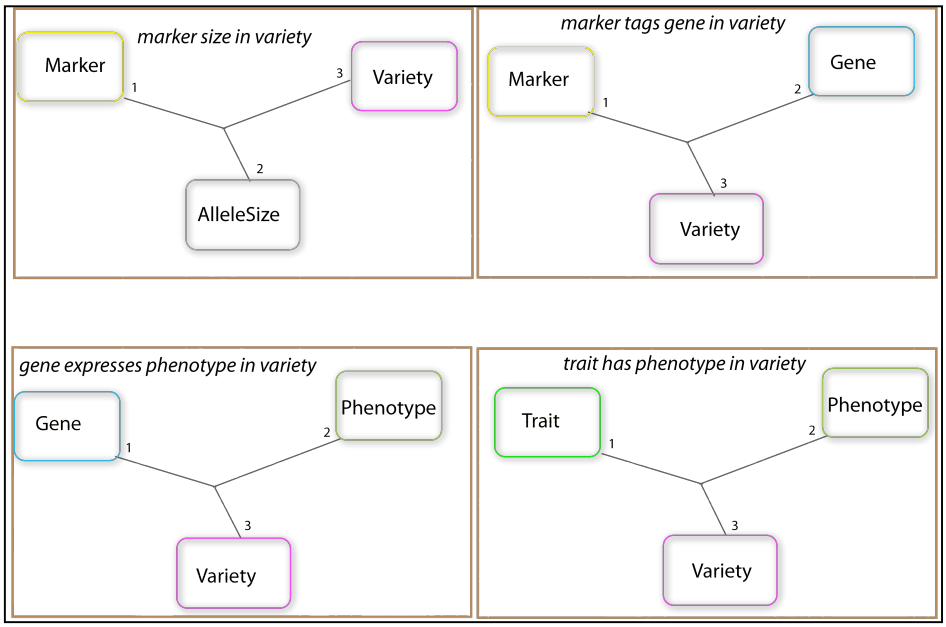


Figure 2. Modèle de connaissances pour la sélection assistée par marqueur génétique.

La méthode automatique d'extraction d'information utilise une analyse profonde du texte et l'apprentissage automatique de la chaîne AlvisNLP [Nedellec et al., 2009]. L'apprentissage automatique supervisé s'applique à des exemples annotés par 13 experts du domaine, principalement des sélectionneurs de semences et les deux auteurs de l'article, biologistes du GDEC. Le corpus annoté est composé de 72 articles sélectionnés pour leur intérêt et leur représentativité et principalement publiés par le journal *Theoretical and Applied Genetics, International Journal of Plant Breeding Research*. Ce journal a été choisi car il mentionne un grand nombre de marqueurs et il est accessible en ligne. Chaque document est annoté deux fois, en double aveugle. Les annotateurs utilisent l'éditeur d'annotation AlvisAE (Figure 3) auquel ils ont été formés [Papazian et al., 2012]. Le document de consignes décrit le modèle, définit les entités et relations et donne de nombreux exemples [Nédellec et al., 2013].

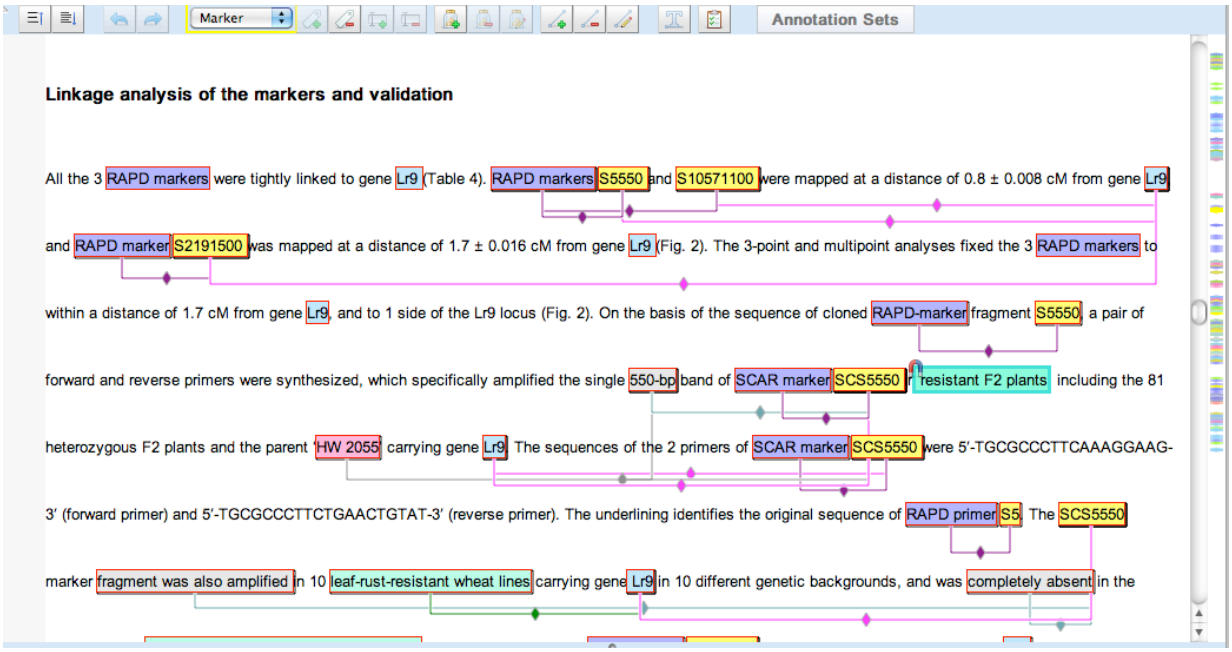


Figure 3. L'éditeur d'annotation AlvisAE.

AlvisAE est d'utilisation intuitive et prend en compte des préannotations automatiques. Les préannotations font gagner un temps précieux, en particulier pour les annotations d'entités très fréquentes et faciles à identifier automatiquement. AlvisAE permet d'annoter des entités discontinues et partiellement recouvrantes, fréquentes dans les textes du corpus. Il permet également d'annoter les coréférences, ce qui évite aux annotateurs de répéter l'annotation de relations redondantes. Seules les sections pertinentes des articles sont annotées complètement. L'annotation en double-aveugle est suivie d'une étape de détection et de résolution de contradictions assistée par AlvisAE. La table 4 montre la répartition des annotations des entités et relations sur 293 sections dans les 72 articles, à ce stade de l'annotation. La distribution est le reflet de l'importance des informations pour les sélectionneurs. Les entités de types gène, variété, trait, marqueur, et méthode sont les plus fréquentes. L'information sur les allèles est rarement explicitée avec le nom de l'allèle, ce qui explique sa faible représentation. Les relations impliquant explicitement les allèles sont également les moins fréquentes. Cette faible représentation n'est pas critique dans la mesure où le nom de l'allèle n'est pas nécessaire pour déterminer le lien entre marqueur et phénotype.

Entités		**Relations binaires**		**Relations ternaires**	
Alelle	368	marker_of_type	307	marker_tags_gene_in_variety	24
AlleleSize	153	marker_tags_gene	184	gene_expresses_phenotype_in_variety	103
Gene	1826	gene_controls_trait	260	trait has phenotype in variety	24
Marker	703	variety_has_phenotype	224	marker_size_in_variety	58
Phenotype	403	allele_expresses_phenotype	107		**207**
Trait	603	is_allele_of_gene	107		
Méthode	508	is_allele_in_variety	64		
Variety	1284	trait has value	34		
	5 848	marker_alleleSize	55		
			1 342		

Table 1. Annotations manuelles des connaissances de marqueurs génétiques.

4 Reconnaissance des entités nommées

Pour concevoir les méthodes de reconnaissance automatique des entités, nous distinguons les noms d'entités figés [Kripke, 1982], les noms de gènes, de marqueurs, de méthodes de d'identification de marqueurs (type de marqueur), de variétés, des noms des entités non figés, les phénotypes et les traits. La reconnaissance des entités figées est réalisée à l'aide de nomenclatures telle que celles de Maswheat[2] et de patrons d'extraction définis manuellement. Les patrons permettent de reconnaître des variations et de désambigüiser à l'aide du contexte. C'est particulièrement pertinent pour les variétés dont les noms ont souvent des homonymes dans les textes comme *Leeds*, qui est à la fois cité comme variété et comme université anglaise. Le taux de reconnaissance calculé en comparant les entités prédites aux entités annotées figure dans le tableau 2.

Les noms de gènes et de marqueurs sont globalement plutôt bien reconnus, mais pas toujours distingués les uns des autres en raison de la proximité de la graphie de leurs noms. Leur annotation manuelle étant globalement de bonne qualité, la source de l'erreur est à chercher dans la méthode. En particulier, un gain est à trouver dans la reconnaissance des bornes des entités de type gène. L'accroissement de 12 points entre l'appariement exact et l'appariement partiel montre que les bornes ne sont pas toujours bien identifiées.

Il est difficile de tirer une conclusion du taux médiocre de reconnaissance des longueurs des allèles. Leurs noms sont en principe réguliers, sous la forme d'un nombre suivi de *bp* (*base pair*) comme dans *103 bp*. Un nombre élevé d'entités de type allèle sont annotées par erreur comme longueur d'allèle, comme *Ppd-D1a*. De nombreux exemple signifient une absence plutôt qu'une longueur, comme par exemple, *absence of PCR products*. La correction rigoureuse de ces annotations permettra d'obtenir une évaluation plus significative.

Le taux de reconnaissance de type de marqueur est étonnamment bas étant donné le nombre réduit de noms de méthodes (par exemple, *AFLP*, *CAPS*, *microsatellite*). Une correction des patrons en fonction des erreurs de prédiction devrait à cours terme augmenter significativement la performance de la méthode.

	Comparaison exacte			**Recouvrement partiel**		
	Rappel	Précision	F1	Rappel	Précision	F1
Gene	0,61	0,49	0,54	0,73	0,61	0,66
Marker	0,58	0,65	0,61	0,59	0,66	0,62
Type	0,54	0,62	0,58	0,56	0,64	0,60
AlleleSize	0,39	0,49	0,43	0,46	0,50	0,48

Table 2. Taux de reconnaissance des entités figées.

L'analyse de l'ensemble des prédictions des entités figées permet donc de faire l'hypothèse que la correction des annotations manuelles erronées, qui est en cours et un ajustement des patrons de reconnaissance devraient significativement accroître la qualité des prédictions.

Pour identifier les entités non figées, traits et phénotypes du blé, et pour l'interrogation de la base de données et du moteur de recherche, nous avons construit la termino-ontologie *WheatPhenotypes* de 624 concepts et synonymes. Elle est organisée hiérarchiquement avec une profondeur maximale de neuf niveaux. Elle comprend deux principaux sous-arbres décrivant les facteurs environnementaux, biotiques et abiotiques et les propriétés de la plante. Les facteurs biotiques sont les facteurs environnementaux qui impliquent des organismes vivants et qui ont un impact sur les caractéristiques du blé. Les facteurs abiotiques décrivent les conditions physiques de développement, (eau, température, vent), mais aussi la composition du sol. Les propriétés de la plante sont organisées en six parties qui sont, la réponse aux facteurs environnementaux, le développement, la reproduction, la transformation du produit, la qualité du produit, fibre et alimentaire. La réponse aux facteurs biotiques, en particulier pathogènes, maladies infectieuses et fongiques est la plus importante. Mais tous les phénotypes sont considérés, aussi bien les phénotypes portant sur la germination et le développement que les phénotypes portant sur les qualités boulangères, au niveau moléculaire

[2] http://maswheat.ucdavis.edu/

et sensoriel (Figure 3). Les terminologies ou ontologies existantes décrivant des phénotypes ne conviennent pas à notre besoin. Elles sont dédiées à des espèces différentes comme HPO[3] (Human Phenotype Ontology) pour l'homme ou ATOL[4] (Animal Trait Ontology for Livestock) pour l'animal, ou trop préliminaires et centrée sur des descriptions physiques et biochimiques comme SPTO[5] (Solanaceae Phenotype Ontology).

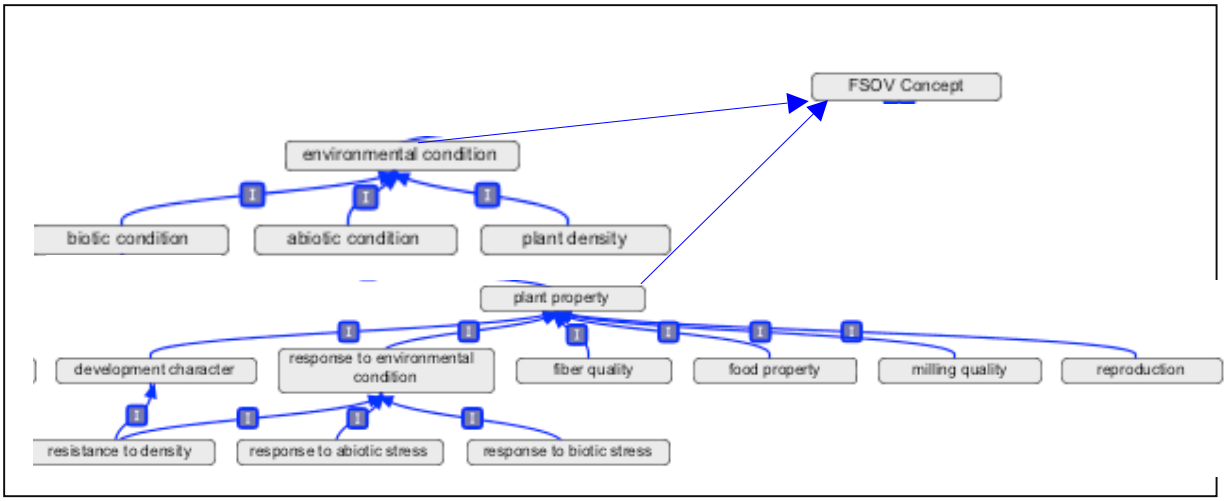


Figure 3. Extrait de l'ontologie *WheatPhenotypes*.

La projection des termes de *WheatPhenotypes* sur les textes, pour identifier les traits et phénotypes, est une méthode d'une grande précision mais d'un rappel médiocre, en raison de la variabilité des termes et de la mention de phénotypes absents de l'ontologie. Pour augmenter le potentiel de reconnaissance des entités à l'aide de l'ontologie, nous utilisons la méthode ToMap [Golik et al., 2011]. Pour prédire des entités, la méthode utilise une terminologie du domaine et les sorties d'un extracteur de termes appliqué sur la collection complète. Son principe est proche de celui de MetaMap pour la reconnaissance de termes UMLS [Aronson et Lang, 2010], mais elle est applicable à toute terminologie, hiérarchique ou non. Elle a montré en particulier son efficacité dans le challenge *BioNLP Shared Task Bacteria Biotope* pour la reconnaissance d'entités de biotopes bactériens [Ratkovic et al, 2012 ; Bossy et al., 2013]. L'extracteur de termes utilisé ici est BioYateA. Il est particulièrement approprié à l'extraction de termes comportant des attachements prépositionnels comme l'attachement *to crown rot* dans *partial seedling resistance to crown rot* [Golik et al., 2013].

Nous avons fait le choix d'évaluer les performances de l'approche en validant manuellement les prédictions faites sur un ensemble de résumés. Les annotations manuelles des phénotypes et traits ne sont pas des références fiables, la phase de consolidation et de validation est encore en cours. Par exemple, le terme *wild type* qui indique une variété non génétiquement modifiée est fréquemment annotée comme un phénotype. *Winter* est annoté comme phénotype au lieu de *winter habit*, qui désigne la période de croissance de la variété. Le phénotype est parfois confondu avec le trait comme *ToxA sensitivity* et *ToxA sensitive*.

Pour obtenir une évaluation plus réaliste nous avons validé manuellement les prédictions des entités par la méthode sur les résumés d'un sous-ensemble de 870 articles. Cette validation permet de mesurer la précision de la prédiction, mais pas son rappel pour lequel une annotation de référence est nécessaire. La table 3 détaille les résultats obtenus. Les premières lignes concernent les termes de phénotype ou de trait, correctement extraits. Nous avons distingué les termes trop généraux qui ne sont pas informatifs pour les sélectionneurs, mais qui sont néanmoins utiles pour la structuration de la connaissance, tel que *morphologic traits*. Les erreurs de prédiction sont réparties en erreurs de pré-traitement linguistique, soit de segmentation en mots, soit d'étiquetage grammatical qui est réalisé par TreeTagger [Schmid, 1994], ou erreurs de paramétrage de la méthode, ou erreurs plus fondamentales de la méthode.

[3] http://www.human-phenotype-ontology.org/

[4] http://www.atol-ontology.com/index.php/fr/

[5] https://bioportal.bioontology.org/ontologies/SPTO

Le paramétrage adapté au domaine consiste à réviser manuellement la liste des têtes des termes qui sont générales ou qui sont ambigües par rapport à l'objectif de reconnaissance. Le mot *content* est un exemple, il apparaît dans des termes pertinents comme *Grain Protein Content* ou *reduction in DON content*, mais également dans des termes non pertinents comme *polymorphism information content*.

L'application directe de BioYateA puis ToMap produit une précision très élevée de 81 %. Dans 11% des cas d'erreur, le mauvais paramétrage en est la cause. Un examen des termes extraits a montré un nombre important de têtes ambigües pour lesquelles nous avons développé des règles de déambigüisation et de révision des bornes. La précision a augmenté significativement de 14 points atteignant 95 %. Le nombre de termes reconnus a parallèlement diminué de 13 %, passant de 245 à 212 termes. Une partie des 33 termes supprimés a été fusionnée avec d'autres termes, comme *main growth habits* et *growth habits* en conséquence de la révision des bornes. Seule une mesure de rappel sur des annotations de référence permettra dans le futur de mesurer précisément s'il y a perte de rappel.

	Sans désambiguisation		**Avec désambiguisation**	
Validation	**Nb termes**	**Proportion**	**Nb de termes**	**Proportion**
Positif	245	81 %	212	95 %
Correct et précis	227	76 %	176	79 %
Correct et général	18	6 %	36	16 %
Négatif	54	19 %	11	5%
Erreur d'analyse linguistique	5	1,7%	4	2 %
Erreur de la méthode	16	5,4%	7	3 %
Erreur de paramètre	33	11 %	0	0 %

Table 3. Précision de la prédiction automatique des phénotypes.

La mesure de précision de la prédiction des entités de type phénotypes et traits est donc de niveau exceptionnel sur un problème difficile, mais elle devra être mise en balance avec la mesure du rappel, une fois consolidées les annotations manuelles du corpus de référence.

5 Conclusion

La masse d'informations disponibles dans la littérature scientifique sur les marqueurs génétiques en fait un enjeu de taille pour la sélection génétique. Dans le cas du blé, cette information est particulièrement critique pour les sélectionneurs. Nous avons proposé une formalisation du problème sous la forme d'un modèle de connaissances riche, associé à une ontologie des phénotypes et facteurs environnementaux qui ont été validés par les sélectionneurs. Nous avons défini un cadre d'annotation d'exemples pour une extraction par apprentissage automatique. Les annotations réalisées permettent d'ores et déjà de mettre en œuvre des méthodes d'extraction automatique adaptées au problème. Nous avons proposé plusieurs approches d'extraction des entités basées sur des stratégies linguistiques dont les résultats préliminaires sont encourageants.

Les résultats sont exploitables dans un moteur de recherche sémantique qui indexe les textes complets de la collection des articles pertinents sur le sujet.

La consolidation des annotations permettra de mesurer plus précisément la qualité des méthodes, de mieux les entraîner et de mettre en œuvre l'extraction des relations. L'ensemble de la démarche sera ensuite généralisable à d'autres plantes d'intérêt agronomique, comme le maïs.

6 Remerciements

Le travail est financé partiellement par Oséo grâce au projet Quaero et par FSOV dans le projet SAM blé. Les auteurs remercient pour leur contribution à l'annotation du corpus Jérôme Auzanneau (Agri-Obtention), Stéphane Boury (Caussade Semences), Emmanuelle

Cariou-Pham (Arvalis), Clément Debiton (Unisigma), Noëmie Desmouceaux (Syngenta), Laure Duchalais (RAGT), Ellen Goudemand (Florimond Desprez), Pierre-Marie Le Roux (Secobra), Vanessa S. Windhausen (Saaten Union), Stephen Sunderwirth (Momont).

Références

Aronson AR, Lang FM: An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association* 2010, 17(3):229-36.

Bossy R, Golik W, Ratkovic Z, Bessières P, and Nédellec C: BioNLP Shared Task 2013 – an overview of the bacteria biotope task. *Proc BioNLP Shared Task 2013 Workshop* 2013, pages 74-82. Association for Computational Linguistics (ACL).

Learning to Recognize Phenotype Candidates in the Auto-Immune Literature Using SVM Re-Ranking

Collier N, Tran M-v, Le H-q, Ha Q-T, Oellrich A, et al. (2013) Learning to Recognize Phenotype Candidates in the Auto-Immune Literature Using SVM Re-Ranking. *PLoS ONE* 8(10): e72965. doi: 10.1371/journal.pone.0072965

Golik W, Warnier P, and Nédellec C: Corpus-based extension of termino-ontology by linguistic analysis: a use case in biomedical event extraction. *Proc 9th Intl Conf Terminology and Artificial Intelligence (TIA 2011)* 2011, pages 37-9.

Golik W, Bossy R, Ratkovic Z, Nédellec C: Improving term extraction with linguistic analysis in the biomedical domain. *Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing'13)*, 24-30 march, Samos, Greece; 2013.

Hirschman L, Yeh A, Blaschke C, Valencia A: Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC Bioinformatics* 2005, 6(Suppl 1):S1.

Kim JD, Ohta T, Pyysalo S, Kano Y, Tsujii J: Extracting bio-molecular events from literature – The BioNLP'09 Shared Task. *Computational Intelligence* 2011, 27(4): 513-40.

Kripke, Saul. *Naming and Necessity*. Boston: Harvard University Press, 1982.

Nédellec C., Nazarenko A. et Bossy R., "Information Extraction", Ontology Handbook., S. Staab, R. Studer (eds.), Springer Verlag, Berlin (DEU) : Springer Science - Business Media Deutschland GmbH (International Handbooks on Information Systems), 2nde édition révisée, pages 663-686, 2009.

Nédellec C: Learning Language in Logic – Genic Interaction Extraction Challenge. *Proc 4th Learning Language in Logic Workshop (LLL05)* 2005, pages 31-7.

Nédellec C., Bossy R., Ranoux M., Valsamou D., Sourdille P., *Consignes d'annotation d'articles sur la sélection du blé par marqueurs génétiques*. Projet FSOV Sam Blé. avril 2013.

Papazian F, Bossy R, Nédellec C: AlvisAE: a collaborative Web text annotation editor for knowledge acquisition. *Proc 6th Linguistic Annotation Workshop (The LAW VI)* 2012, pages 149-52.

Paux E, Faure S, Choulet F, Roger D, Gauthier V, Martinant J-P, Sourdille P, Balfourier F, Lepaslier M-C, Brunel D, Cakir M, Gandon B, Feuillet C (2009) Insertion site based polymorphism markers open new perspectives for genome saturation and marker-assisted selection in wheat. Plant Biotechnol J.

Ratkovic Z, Golik W, Warnier P: Event extraction of bacteria biotopes: a knowledge-intensive NLP-based approach. *BMC Bioinformatics* 2012, 13(Suppl 11):S8.

Raats D, Frenkel Z, Krugman T, Dodek I, Sela H, Simková H, Magni F, Cattonaro F, Vautrin S, Bergès H, Wicker T, Keller B, Leroy P, Philippe R, Paux E, Doležel J, Feuillet C, Korol A, Fahima T. The physical map of wheat chromosome 1BS provides insights into its gene space organization and evolution. *Genome Biol*. 2013 Dec 20;14(12):R138. PubMed PMID: 24359668.

Helmut Schmid. Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of International Conference on New Methods in Language Processing, Manchester, UK, 1994.