Caractérisation de souches bactériennes à l'aide de la logique propositionnelle

Fabien Chhel Frédéric Lardeux Frédéric Saubion

LERIA, Université d'Angers
{nom}@info.univ-angers.fr

Ces dernières années ont vu le développement d'importantes collaborations entre les biologistes et les informaticiens qui ont ainsi mis en lumière de nouveaux problèmes fondamentaux, en particulier dans le domaine de l'optimisation combinatoire, comme le problème d'alignement de séquences ou de reconstruction phylogénétique, et dont la résolution requière la conception d'algorithmes performants. En effet, dès lors que l'acquisition de données a fait d'énormes progrès, notamment en terme de séquençage ou d'utilisation de puces à ADN, les volumes de données expérimentales et de mesures ont été considérablement accrus et leur traitement effectif nécessite le développement de nouveaux outils et techniques. Soulignons d'ailleurs au passage qu'une utilisation pratique de ces données se fonde souvent sur une chaîne de compétences scientifiques allant du traitement d'images ou du signal à l'algorithmique en passant par l'utilisation d'outils statistiques d'analyse de données.

Si nous ne nous plaçons pas ici directement dans le champs académique, désormais bien reconnu, de la bioinformatique, l'objectif de cet article est de présenter une application des techniques de modélisation et de résolution en logique propositionnelle dans le domaine de la biologie végétale. Le problème de caractérisation que nous abordons pourra sans nul doute trouver d'autres applications, dont nous avons déjà identifié certaines, en particulier dans le domaine du diagnostic biologique.

La caractérisation précise de collections de souches bactériennes représente un enjeu scientifique majeur puisque les bactéries phytopahogènes sont en effet responsables d'importants dégâts sur les cultures et certaines sont recensées sur des listes d'organismes de quarantaine et font l'objet d'une lutte officielle (Directive 2000/29/CE). Le développement de tests de diagnostic permettant d'identifier des souches de ces espèces s'avère dès lors nécessaire.

Dans ce contexte nous nous intéressons aux souches bactériennes du genre *Xanthomonas*. La notion de pathovar est une subdivision de l'espèce bactérienne phytopathogène qui regroupes des souches responsables d'un même symptôme sur une espèce végétale ou une gamme d'espèces végétales. En particulier les Xanthomonas sont un modèle d'études puisqu'elles présentent une centaine de pathovars différents. Par exemple, le *Xanthmonas axonopodis* se décline en *pathovar citri* qui occasionne le chancre citrique des agrumes mais également en *pathovar vesicatoria* qui est responsable de la gale bactérienne du poivron. Toutefois, la phylogénie des souches ne suffit pas à expliquer la spécifié d'hôte, c'est à dire l'espèce végétale attaquée par la souche. En particulier, certains pathovars proches génétiquement peuvent avoir des spécificités d'hôtes très éloignées et vice versa. L'approche consiste alors à identifier pour les souches des répertoires de gènes pertinents (gènes de virulence) et à analyser la corrélation entre la présence/absence de ces gènes et la spécificité d'hôtes des pathovars (groupes de souches bactériennes) [3].

Récemment, la description de répertoires de 35 gènes de virulence dans une collection de 132 souches de Xanthomonas réparties dans 21 groupes et présentant des spécificités d'hôte différentes a permis de montrer qu'il existe une corrélation entre le répertoire de gènes de virulence d'une souche de Xanthomonas et sa spécificité d'hôte. Au sein du genre Xanthomonas, d'autres espèces sont inscrites sur les listes de quarantaine, ou même sur les listes de bioterrorisme, et font ainsi l'objet de réglementations strictes.

Le problème de caractérisation se pose alors comme l'identification d'une famille de souches par rapport aux autres familles en fonction de la présence ou de l'absence de certains gènes. Une souche sera donc un vecteur de valeurs binaires qui rendent compte de la présence (valeur 1) ou de l'absence (valeur 0) d'un caractère.

Plus concrètement, une instance de problème possédant 5 souches, réparties en 3 groupes et basée sur un ensemble de 4 gènes peut être illustrée par la figure .

Résoudre ce problème revient à caractériser chaque

groupe. Il faut donc, pour chaque groupe, trouver une combinaison de présences ou d'absences de gènes valide pour toutes les souches du groupe et non valide pour toutes les autres souches des autres groupes. Dans l'exemple de la figure , le groupe 1 est caractérisé par la présence conjointe des gènes grisés.

Souche	Groupe	Gènes			
		x1	x2	x3	x4
e1	g1	1	1	1	0
e2	g1	1	1	1	1
e3	g2	0	0	1	0
e4	g2	0	1	1	1
e5	g3	1	1	0	0

FIGURE 1 – Exemple de caractérisation

Une fois posé le problème de la caractérisation, naturellement diverses méthodes peuvent être envisagées. En particulier, la méthode des CCD (Coefficient de Capacité de Diagnostique [1]), dans laquelle, par le biais d'une étude statistique, les gènes sont triés par pertinence de caractérisation. L'atout d'une telle approche est la simplicité du calcul mais elle ne permet de traiter que la caractérisation sur un seul gène. Il existe alors un réel besoin de développer de nouvelles approches pour fournir des formules de caractérisation combinant plusieurs gènes. De plus, les biologistes sont intéressés par deux propriétés spécifiques des solutions :

- Une solution qui minimise le nombre de caractères utilisés. Ceci est d'autant plus important que, dans l'optique de la fabrication de tests de diagnostic basé sur des puces à ADN [5], le nombre de gènes à observer doit être minimisé à la fois pour des raisons de coûts, de temps d'expérience et de fiabilité. Enfin, un autre critère lié à la présence plutôt que leur absence des gènes peut intervenir, car il est plus facile d'observer expérimentalement l'existence d'un gène à l'aide de marqueurs que de vérifier son absence.
- Le calcul de toutes les solutions : il peut s'avèrer utile, du point de vue de l'interprétation biologique, de disposer d'une représentation de toutes les solutions possibles car elles pourraient faire apparaître des relations particulières entre les gènes permettant d'expliquer certaines spécificités fonctionnelles des bactéries.

D'un point de vue formel, nous pouvons considérer que les absences ou présences de gènes peuvent être vue comme les valeurs de vérités de variables booléennes et que la caractérisation d'un groupe revient alors à trouver une formule booléenne qui est satisfaite par les affectations correspondant à ce groupe et falsifiée par les autres. La relation avec l'apprentissage de fonction booléenne surgit alors naturellement. L'apprentissage automatique de fonction booléenne à partir d'exemples est un problème qui a été très largement étudié depuis de nombreuses années, depuis les travaux initiaux de Valiant [7] puis de Natarajan [4], jusqu'à des travaux plus récents comme Gavalda et al. [2]. Toutefois, le problème est ici différent. La caractérisation doit être faite pour chaque groupe par rapport aux autres de manière exacte. Enfin, l'objectif est de minimiser le nombre de littéraux utilisées pour cette caractérisation croisée. L'aspect combinatoire de cette caractérisation de chaque groupe par rapport aux autres induit d'ailleurs une complexité importante du problème.

Dans ce travail, nous modélisons problème de caractérisation comme la recherche d'ensembles de formules en logique propositionnelle, ce qui nous permet ensuite de mettre en évidence que sa classe de complexité est Σ_2^p-complet, puisqu'il revient à minimiser des formules DNF [6]. Nous avons développé un algorithme de caractérisation incluant des techniques de simplification qui nous a permis d'obtenir des résultats, ayant déjà été valorisés par la mise au point de tests de diagnostic collaboration avec nos collègues de l'UMR PAVE de l'INRA d'Angers.

Références

[1] P. Descamps and M. Véron. Une méthode de choix des caractères d'identification basée sur le théorème de bayes et la mesure de l'information. *Ann. Microbiol. (Paris)*, 132B, 1981.

[2] Ricard Gavaldà and Denis Thérien. An algebraic perspective on boolean function learning. In *Algorithmic Learning Theory, 20th International Conference, ALT 2009, Porto, Portugal, October 3-5, 2009. Proceedings*, volume 5809 of *Lecture Notes in Computer Science*, pages 201–215. Springer, 2009.

[3] Ahmed Hajri, Chrystelle Brin, Gilles Hunault, Frédéric Lardeux, Christophe Lemaire, Charles Manceau, Tristan Boureau, and Stéphane Poussier. A "repertoire for repertoire" hypothesis : Repertoires of type three effectors are candidate determinants of host specificity in xanthomonas. *PLoS ONE*, 4(8) :e6632, 08 2009.

[4] B. K. Natarajan. On learning boolean functions. In *Proceedings of the Nineteenth Annual ACM Symposium on Theory of Computing*, pages 296–304. ACM, 1987.

[5] M. Schena, D. Shalon, R.W. Davis, and P.O. Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235) :467–470., 2005.

[6] Christopher Umans. The minimum equivalent DNF problem and shortest implicants. *J. Comput. Syst. Sci.*, 63(4) :597–611, 2001.

[7] L. G. Valiant. A theory of the learnable. *Commun. ACM*, 27(11) :1134–1142, 1984.