OPINION PAPER

Towards predictive resistance models for agrochemicals by combining chemical and protein similarity via proteochemometric modelling

**Gerard J. P. van Westen · Andreas Bender ·
John P. Overington**

Abstract Resistance to pesticides is an increasing problem in agriculture. Despite practices such as phased use and cycling of 'orthogonally resistant' agents, resistance remains a major risk to national and global food security. To combat this problem, there is a need for both new approaches for pesticide design, as well as for novel chemical entities themselves. As summarized in this opinion article, a technique termed 'proteochemometric modelling' (PCM), from the field of chemoinformatics, could aid in the quantification and prediction of resistance that acts via point mutations in the target proteins of an agent. The technique combines information from both the chemical and biological domain to generate bioactivity models across large numbers of ligands as well as protein targets. PCM has previously been validated in prospective, experimental work in the medicinal chemistry area, and it draws on the growing amount of bioactivity information available in the public domain. Here, two potential applications of proteochemometric modelling to agrochemical data are described, based on previously published examples from the medicinal chemistry literature.

G. J. P. van Westen (✉) · J. P. Overington
European Molecular Biology Laboratory European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom
e-mail: gerardvw@ebi.ac.uk

A. Bender
Unilever Centre for Molecular Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, United Kingdom

Introduction

In agriculture, resistance to pesticides forms a complex and growing problem, which includes the development of resistance to insecticides [17], fungicides [16], as well as herbicides [7, 25, 6]. For each of these resistance types, a multitude of different resistance mechanisms are possible, all of which, however, lead to phenotypic resistance (i.e. the concentration of pesticide needed to kill the pests is higher for resistant variants compared to wild type). Commonly observed resistance mechanisms are similar to those observed in microbial and cancer mechanisms of resistance; examples include increased expression of efflux proteins, increased expression of metabolizing proteins, and point mutations in the protein targeted by the agrochemical agent [8, 10, 24]. Due to the spectrum of possible adaptations in the target organism, it is difficult to capture and model all potential forms of resistance for a certain compound in a model a priori, which is analogous to antibiotic and anti-cancer drug resistance. Out of these possibilities, the current opinion article will deal specifically with the impact of point mutations at the ligand-binding site and their effect on resistance. This is an area for which there is prior successful experience in the medicinal chemistry and drug design field, including prospective experimental validation of the models developed. Here, previous research of the authors as well as other related groups will be outlined, with the aim to transfer these methods also to the world of agrochemical research [22, 20].

Complementary ligand and target information

The binding affinity between a ligand (e.g. a small molecule or RNAi) [26, 1], and a target (usually a protein, but potentially a ribosomal target) [18], is governed by properties of both ligand and target (and the physiological environment, e.g. ionic

strength and pH). Hence, it follows that changes to either one of the binding partners will, in most cases, translate into a difference (weaker or stronger binding) in affinity and subsequent efficacy. This mechanism is also exploited when a candidate molecule is optimized through iterative design of synthetic molecules to have a high affinity towards a desired protein target via quantitative structure-activity relationship (QSAR) studies, which is a standard technique during lead optimization in drug discovery. In the context of agrochemistry, extending this principle from single targets to multiple targets would be equivalent to optimizing a candidate molecule towards the desired pest, while avoiding unwanted effects, e.g. in humans (as well as other species). This procedure exploits the differences in the sequences and structures between the targets in these species (corresponding to a variation on the target side). In order to now model the activity of multiple ligands against multiple targets, 'proteochemometric modelling' (PCM) can be used, a computational technique that simultaneously uses properties of both ligand and target space. Hence, PCM can be used to make predictions for unknown ligands as well as unknown protein targets (Fig. 1; for reviews see [15, 23]). For clarity, these new sequences could be an orthologous sequence from a new pathogen or new sequences generated under selective pressure giving rise to resistant pathogens. The reader is referred to recent reviews and references therein for an overview of the more general concept of computational models of bioactivity and (QSAR) models (these include an overview of

various applicable machine learning algorithms and descriptors) [23, 5]. Previously, PCM has been successfully applied to model the impact of protein mutations on the resistance of viruses to drugs. Examples include the human immunodeficiency virus (HIV), also including prospective experimental validation [14, 22, 20], as well as the dengue virus [19]. Given that the problem at hand is similar—in both cases, the impact of sequence mutations/differences of a target protein on ligand binding or efficacy needs to be understood—we anticipate that the concept of PCM will also be transferable to the field of agrochemicals, possibly with some domain-dependent modifications.

Resistance models

The first aim of PCM is to model the bioactivity of ligands against large number of related proteins, which allows the scientist to anticipate the bioactivity spectrum of a molecule in a multidimensional fashion. In the literature, it has been shown that multitarget models combining ligand and target properties (i.e. PCM models) generally perform better than single target models, which are trained on ligand-only properties (i.e. conventional QSAR models) [14, 19, 22, 23, 20]. Moreover, the multitarget nature of these models allows rationalization of resistance, as these models were able to deconvolute the individual contributions of amino acid substitutions (Fig. 2a [20]) or substructures within ligands that
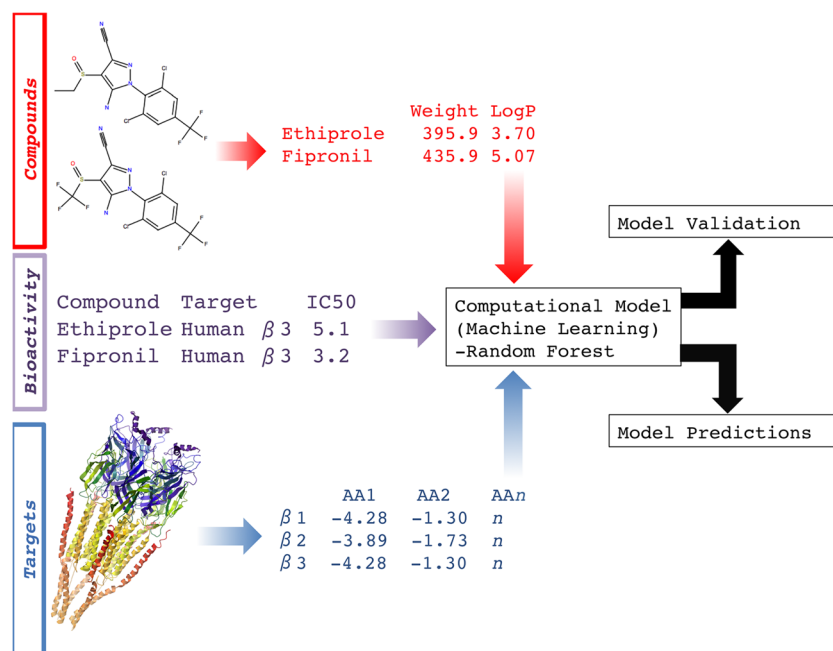

Fig. 1 Schematic overview of proteochemometric modelling (PCM). The technique uses chemical properties from multiple ligands (e.g. ethiprole), target properties from multiple proteins (e.g. gamma-aminobutyric acid (GABA) receptor subunit beta-3 [4]), and their respective bioactivity values. From these data, a statistical model is generated. After validation, this model can predict the bioactivity of untested compounds on the targets included in the model, in order to (1) select a compound active on a particular protein target; (2) predict which protein a particular compound will show activity against; or (3) for the prospective selection ('virtual screening') of compounds with a desired bioactivity profile against a set of proteins and its mutants
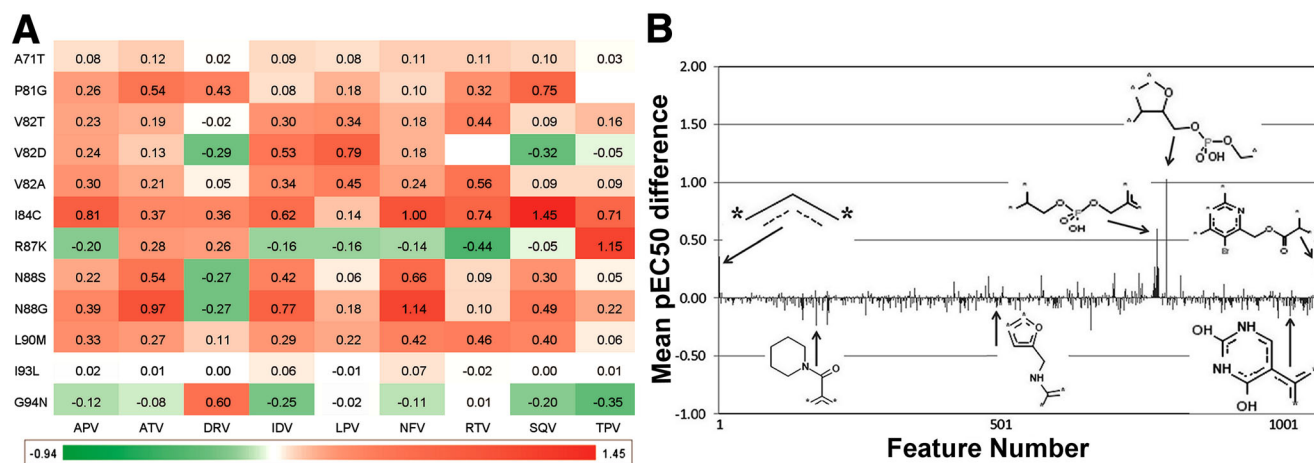
A

	APV	ATV	DRV	IDV	LPV	NFV	RTV	SQV	TPV
A71T	0.08	0.12	0.02	0.09	0.08	0.11	0.11	0.10	0.03
P81G	0.26	0.54	0.43	0.08	0.18	0.10	0.32	0.75	
V82T	0.23	0.19	-0.02	0.30	0.34	0.18	0.44	0.09	0.16
V82D	0.24	0.13	-0.29	0.53	0.79	0.18		-0.32	-0.05
V82A	0.30	0.21	0.05	0.34	0.45	0.24	0.56	0.09	0.09
I84C	0.81	0.37	0.36	0.62	0.14	1.00	0.74	1.45	0.71
R87K	-0.20	0.28	0.26	-0.16	-0.16	-0.14	-0.44	-0.05	1.15
N88S	0.22	0.54	-0.27	0.42	0.06	0.66	0.09	0.30	0.05
N88G	0.39	0.97	-0.27	0.77	0.18	1.14	0.10	0.49	0.22
L90M	0.33	0.27	0.11	0.29	0.22	0.42	0.46	0.40	0.06
I93L	0.02	0.01	0.00	0.06	-0.01	0.07	-0.02	0.00	0.01
G94N	-0.12	-0.08	0.60	-0.25	-0.02	-0.11	0.01	-0.20	-0.35

-0.94 ▬▬▬▬▬▬ 1.45

B

Fig. 2 Proteochemometric models can be interpreted biologically (**a**) and chemically (**b**). **a** Displays the average contribution of individual amino acid substitutions to resistance measured as a fold change. Fields are *coloured green* when the combination of mutation and inhibitor is more sensitive, and *red* when the combination is resistant. The *columns* represent clinical HIV protease inhibitors; the *rows* represent point mutations in the following notion: wild-type residue, residue position, and mutated residue (e.g. V82T, valine to threonine at position 82). It can be observed that individual mutations can have different effects on different inhibitors, many of which could be related to experiment (see main text for details). **b** Shows the average contribution of chemical substructures to pEC$_{50}$. The *bars* are positive if the respective feature contributes positively to bioactivity, and negative where the feature leads to lower activity. The *y-axis* indicates the average contribution over all present ligands (451) and mutants (14). This information can be used to guide compound selection as well as optimization in cases where bioactivity against multiple protein targets needs to be taken into account. (Fig. adapted from [20] and [22])

vary across a chemical series (Fig. 2b [22]). For instance in the case of HIV protease, it is known that residue 82 can mutate from the wild-type valine to a threonine, and that this simple, conservative, single-point mutation leads to (cross) resistance amongst all clinical inhibitors with the exception of Darunavir [12]. PCM models have been able to reproduce these results, as shown in Fig. 2a.

In the case of agrochemistry, the authors are of the opinion that the nature of the PCM technique would be equally suited to identify potential agrochemicals that have the most favourable resistance profile. Similarly, models could be used to deconvolute contributions of mutations to an increase or reduction of resistance displayed by individual mutants.

Multispecies models

Knowledge on the relationship between homologues targets in different species can be exploited using PCM, which is of relevance either in cases where off-target effects of a compound in a species needs to be avoided, but also importantly where the aim is to target multiple distinct species in a designed spectrum of activity. Previously, a proof-of-concept study [21] employed a single bioactivity model to simultaneously model both the human and rat orthologues of the adenosine receptors (G protein-coupled receptors) using data from the ChEMBL [9] database. The model was able to identify several novel ligands that were experimentally validated,

and one of the ligands showed high affinity in the nanomolar range. Upon further inspection, it could be found that the selection of this ligand from a database was likely due to information from other species, underlining the value of integrating as much information from bioactivity space as reasonably possible. In Fig. 3a, a multidimensional scaling analysis (MDS) of the similarity between these eight proteins is shown. From the figure, it is apparent that orthologues (genes closely related in sequence and having the same function in different species) are more similar in the particular definition used than paralogues (genes which are similar in sequence but which have different functions in the same species). Constructing multispecies models on this type of data allows the rationalization of differences in activity between different species, as well as its application for compound design (such as in the above study). As is the case in the application to resistance, this model interpretation can be performed both from the ligand (chemical) point of view and from the target (protein or RNA) point of view, leading on the one hand to the elucidation of chemical features to guide compound optimization, and on the other hand to mutations driving resistance from the protein side.

These approaches are also transferable to the pesticide field. Analogous to the aforementioned adenosine receptors, gamma-aminobutyric acid A (GABA-A) ligand-gated ion channels can be aligned to capture and represent the (dis)similarities between species. These ion channels form the target for phenylpyrazole insecticides, and resistance has been demonstrated through point mutations
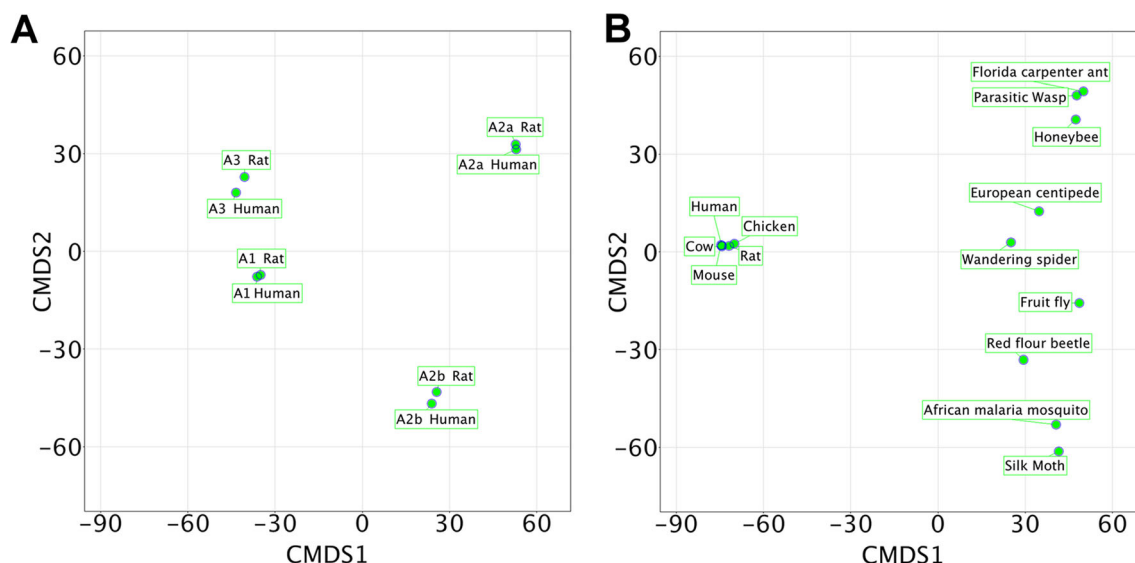
Fig. 3 Multidimensional scaling (MDS) plots demonstrating the similarity between protein families in different species. The further points are located from each other, the more dissimilar they are. **a** Full protein similarity between the different adenosine receptor isoforms (orthologues and paralogues) for human and rat. Orthologues (identical function genes in different species) are more similar than paralogues (different function but sequence related genes in the same species). **b** Full similarity between the GABA-A ligand-gated ion channels in mammals and a selection of arthropods. The *plot* illustrates that mammallian receptors are more

similar to each other than they are to their insect counterparts, and that insects amongst themselves display a large variation. As it was previously demonstrated that it is possible to model the protein similarity space shown in (**a**) it stands to reason that it is feasible to model the space shown in (**b**). Information of this type can be used (and has been used [21]) in PCM modelling to anticipate, and predict, bioactivity of ligands against orthologues of a protein from multiple species, which is useful for, e.g. off-target modelling

[2, 4, 3]. An MDS of the similarity between mammalian and insect isoforms of these complexes is shown in Fig. 3b, illustrating that mammalian channels are more similar to each other than they are to their insect counterparts. Furthermore, the selected arthropods display a larger variation between species than do the selected mammals. As it was previously demonstrated that it is possible to model the protein similarity space shown in Fig. 3a (for the adenosine receptors), it stands to reason that it is feasible to model the space shown in Fig. 3b (representing the GABA-A ion channels). PCM models are agnostic of the particular target and application area they are used in—their applicability depends on the amount of data available both from the chemical and biological side (however, this requirement should not be neglected). Hence, the data visualized in Fig. 3b should allow for the construction of a predictive model that can predict activity (and toxicity) of candidate compounds on GABA-A in the species included in the analysis, in a manner similar to the adenosine receptors described above.

Finally, it should be noted that the ribosome has gained significant attention as a druggable target (specifically for antibiotics) [18]. It has been shown that bacteria can gain resistance via multiple diverse mutations in 23S ribosomal RNA [11, 13]. Yet a complementary mode of action leading to novel pesticides should provide a useful addition to

established modes of action and increase the resistance threshold.

Conclusions

In summary, PCM is a versatile quantitative modelling technique for interactions between ligands and their biological targets. In the current opinion paper, we highlight application of the technique and precedence for success from related fields and sketch applications to the agrochemical context, comprising insecticides, fungicides, and herbicides. Based on the prospective experimental validation that has been performed for enzymes and receptors in previous studies, such work is likely to be successful. Firstly, there is the modelling and prediction of resistance towards pesticides by weeds, fungi, or insects. A second application is the prediction of activity of pesticides or other agricultural chemicals in organisms were this is undesired ('off-targets'). Finally, the technique can be used in virtual screening in the identification of potential new agrochemicals, aimed at a higher resistance threshold (broader activity against multiple mutants), or potential agrochemicals anticipated to have less toxic effects on non-pest species. We anticipate a great future for studies in this area, as the cost of sequencing (which directly relates to the generation of protein-side descriptors) continues to drop and the amount of bioactivity data available

increases, both of which increases our ability to generate predictive PCM models considerably.

References

1. Auer C, Frederick R (2009) Crop improvement using small RNAs: applications and predictive ecological risk assessments. Trends Biotechnol 27(11):644–51

2. Bloomquist JR (1996) Ion channels as targets for insecticides. Annu Rev Entomol 41(1):163–90

3. Buckingham SD, Biggin PC, Sattelle BM, Brown LA, Sattelle DB (2005) Insect GABA receptors: splicing, editing, and targeting by antiparasitics and insecticides. Mol Pharmacol 68(4):942–51

4. Caboni P, Sammelson RE, Casida JE (2003) Phenylpyrazole insecticide photochemistry, metabolism, and GABAergic action: ethiprole compared with fipronil. J Agric Food Chem 51(24):7055–61. doi:10.1021/jf030439l

5. Cherkasov A, Muratov EN, Fourches D, Varnek A, Baskin II, Cronin M, Dearden J, Gramatica P, Martin YC, Todeschini R, Consonni V, Kuz'min VE, Cramer R, Benigni R, Yang C, Rathman J, Terfloth L, Gasteiger J, Richard A, Tropsha A (2013) QSAR modeling: where have you been? Where are you going to? J Med Chem. doi:10.1021/jm4004285

6. Devine MD, Shukla A (2000) Altered target sites as a mechanism of herbicide resistance. Crop Prot 19(8):881–9

7. Erickson JM, Rahire M, Rochaix J-D, Mets L (1985) Herbicide resistance and cross-resistance: changes at three distinct sites in the herbicide-binding protein. Science 228(4696):204–7

8. Ffrench-Constant RH, Rocheleau TA, Steichen JC, Chalmers AE (1993) A point mutation in a *Drosophila* GABA receptor confers insecticide resistance. Nature 363(6428):449–51

9. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, Overington JP (2012) ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res 40(D1):D1100–7. doi:10.1093/nar/gkr777

10. Hosie AM, Baylis HA, Buckingham SD, Sattelle DB (1995) Actions of the insecticide fipronil, on dieldrin-sensitive and- resistant GABA receptors of *Drosophila melanogaster*. Br J Pharmacol 115(6):909–12. doi:10.1111/j.1476-5381.1995.tb15896.x

11. Hummel H, Boück A (1987) 23S ribosomal RNA mutations in halobacteria conferring resistance to the anti-80S ribosome targeted antibiotic anisomycin. Nucleic Acids Res 15(6):2431–43

12. Johnson VA, Calvez V, Paredes R, Pillay D, Shafer RW, Wensing AM, Richman DD (2013) Update of the drug resistance mutations in HIV-1: March 2013. Top Antivir Med 21:6–14

13. Kim JM, Kim JS, Kim N, Kim Y-J, Kim IY, Chee YJ, Lee C-H, Jung HC (2008) Gene mutations of 23S rRNA associated with clarithromycin resistance in *Helicobacter pylori* strains isolated from Korean patients. J Microbiol Biotechnol 18(9):1584–9

14. Lapins M, Eklund M, Spjuth O, Prusis P, Wikberg JES (2008) Proteochemometric modeling of HIV protease susceptibility. BMC Bioinforma 9:181. doi:10.1186/1471-2105-9-181

15. Lapinsh M, Prusis P, Gutcaits a, Lundstedt T, Wikberg JE (2001) Development of proteo-chemometrics: a novel technology for the analysis of drug-receptor interactions. Biochim Biophys Acta 1525:180–90

16. Ma Z, Michailides TJ (2005) Advances in understanding molecular mechanisms of fungicide resistance and molecular detection of resistant genotypes in phytopathogenic fungi. Crop Prot 24(10):853–63

17. Oakeshott JG, Devonshire AL, Claudianos C, Sutherland TD, Horne I, Campbell PM, Ollis DL, Russell RJ (2005) Comparing the organophosphorus and carbamate insecticide resistance mutations in cholin- and carboxyl-esterases. Chem Biol Interact 157:269–75

18. Poehlsgaard J, Douthwaite S (2005) The bacterial ribosome as a target for antibiotics. Nat Rev Microbiol 3(11):870–81

19. Prusis P, Lapins M, Yahorava S, Petrovska R, Niyomrattanakit P, Katzenmeier G, Wikberg JES (2008) Proteochemometrics analysis of substrate interactions with dengue virus NS3 proteases. Bioorg Med Chem 16:9369–77. doi:10.1016/j.bmc.2008.08.081

20. Van Westen GJP, Hendriks A, Wegner JK, IJzerman AP, Van Vlijmen HWT, Bender A (2013) Significantly improved HIV inhibitor efficacy prediction employing proteochemometric models generated from antivirogram data. PLoS Comput Biol 9(2):e1002899

21. Van Westen GJP, Van den Hoven OO, Van der Pijl R, Mulder-Krieger T, de Vries H, Wegner JK, IJzerman AP, Van Vlijmen HWT, Bender A (2012) Identifying novel adenosine receptor ligands by simultaneous proteochemometric modeling of rat and human bioactivity data. J Med Chem 55(16):7010–20. doi:10.1021/jm3003069

22. Van Westen GJP, Wegner JK, Geluykens P, Kwanten L, Vereycken I, Peeters A, IJzerman AP, Van Vlijmen HWT, Bender A (2011) Which compound to select in lead optimization? Prospectively validated proteochemometric models guide preclinical development. PLoS ONE 6:e27518. doi:10.1371/journal.pone.0027518

23. Van Westen GJP, Wegner JK, IJzerman AP, Van Vlijmen HWT, Bender A (2011) Proteochemometric modeling as a tool for designing selective compounds and extrapolating to novel targets. Med Chem Commun 2:16–30

24. Walsh C (2003) Antibiotics: actions, origins, resistance. Am Soc Microbiol (ASM)

25. Warwick SI (1991) Herbicide resistance in weedy plants: physiology and population biology. Annu Rev Ecol Evol Syst 22:95–114

26. Whyard S, Singh AD, Wong S (2009) Ingested double-stranded RNAs can act as species-specific insecticides. Insect Biochem Mol Biol 39(11):824–32