OXFORD

Briefings in Bioinformatics, 16(5), 2015, 820–829

doi: 10.1093/bib/bbu040
Advance Access Publication Date: 6 November 2014
Paper

Case study: using sequence homology to identify putative phosphorylation sites in an evolutionarily distant species (honeybee)

Brett Trost, Scott Napper and Anthony Kusalik

Corresponding author. B. Trost, Vaccine and Infectious Disease Organization, University of Saskatchewan, 120 Veterinary Road, Saskatoon, SK, Canada S7N 5E3. Tel: +306 966 1495; Fax: +306 966 7478; E-mail: brett.trost@usask.ca

Abstract

The majority of scientific resources are devoted to studying a relatively small number of model species, meaning that the ability to translate knowledge across species is of considerable importance. Obtaining species-specific knowledge enables targeted investigations of the biology and pathobiology of a particular species, and facilitates comparative analyses. Phosphorylation is the most widespread posttranslational modification in eukaryotes, and although many phosphorylation sites have been experimentally identified for some species, little or no data are available for others. Using the honeybee as a test organism, this case study illustrates the process of using protein sequence homology to identify putative phosphorylation sites in a species of interest using experimentally determined sites from other species. A number of issues associated with this process are examined and discussed. Several databases of experimentally determined phosphorylation sites exist; however, it can be difficult for the nonspecialist to ascertain how their contents compare. Thus, this case study assesses the content and comparability of several phosphorylation site databases. Additional issues examined include the efficacy of homology-based phosphorylation site prediction, the impact of the level of evolutionary relatedness between species in making these predictions, the ability to translate knowledge of phosphorylation sites across large evolutionary distances and the criteria that should be used in selecting probable phosphorylation sites in the species of interest. Although focusing on phosphorylation, the issues discussed here also apply to the homology-based cross-species prediction of other post-translational modifications, as well as to sequence motifs in general.

Key words: phosphorylation; protein kinases; homology; biological databases; honeybee; Apis mellifera

Introduction

Biological research tends to be concentrated on a limited number of model species, mainly those used as representatives for human health (e.g. mice, chimpanzees and humans themselves), those of economic importance (e.g. livestock and food crops) and those with characteristics (such as short generation time) that make them convenient to study (e.g. *Arabidopsis thaliana* and *Drosophila melanogaster*). However, as much valuable work is targeted at non-model species, the ability to translate knowledge across species is of considerable importance. For instance, when a new genome is sequenced, the already characterized genomes of related species are often used to aid the assembly and annotation of the new one.

Protein phosphorylation, which is catalyzed by protein kinases, is the most important posttranslational modification in

Brett Trost is a postdoctoral researcher at the Vaccine and Infectious Disease Organization at the University of Saskatchewan. His research interests include machine learning, sequence analysis and immunoinformatics.
Scott Napper is a senior scientist at the Vaccine and Infectious Disease Organization, and a professor in the Department of Biochemistry, at the University of Saskatchewan. His research interests include kinome analysis, infectious diseases and prion diseases.
Anthony Kusalik is a professor in the Department of Computer Science at the University of Saskatchewan. His research interests cover many subfields of bioinformatics, as well as logic programming.
Submitted: 29 July 2014; **Received (in revised form):** 14 October 2014

eukaryotes. The kinome (the complement of protein kinases in an organism) has been characterized for many species (e.g. yeast [1], human [2] and mouse [3]), and it is straightforward to use sequence homology or structural considerations to translate knowledge of protein kinases among species. Although not as complete as our knowledge of protein kinases, extensive information is available concerning phosphorylation sites (P-sites)—the residues phosphorylated by protein kinases—in many species. For example, the PhosphoSitePlus database [4, 5] currently contains >150 000 human sites and >60 000 mouse sites. The phosphoproteomes of certain plants have also been characterized [6, 7]. Conversely, few, if any, P-sites have been experimentally determined for most species, necessitating the development of techniques that enable the translation of knowledge concerning P-sites across species.

Many machine learning methods for P-site prediction have been proposed [8, 9]. However, most focus on human sites, for which a large amount of experimental data are available. Previously, we published a protocol for using sequence homology to identify putative P-sites in a species of interest (the 'target species') using experimentally determined sites from other species [10]. This protocol involves using BLAST to search for homologs of short peptides in the proteome of the target species. Each such peptide is 15 amino acids in length, is a subsequence of an intact protein and contains a known P-site (usually as its central residue). A peptide length of 15 was chosen because it approximates the region often thought to dictate the specificity of protein kinases, and is consistent with the range of lengths found to be effective for machine learning-based predictors (e.g. [11–13]; see also [9]). If a close match is found, then the residue in the matching peptide that corresponds to the P-site in the query peptide is a probable P-site. DAPPLE (http://saphire.usask.ca/saphire/dapple) is a software pipeline that automates this protocol [14]. DAPPLE takes as input a list of peptides containing known P-sites, as well as the proteomes of the organisms in which those P-sites are found, and outputs several pieces of information concerning each peptide's best match in the target proteome. These include the protein in which the best match was found, the number of sequence differences between the query peptide and its best match, the number of non-conservative sequence differences; the locations of the P-sites in the full proteins and whether the full proteins are reciprocal BLAST hits (and thus likely orthologs). Table 1 gives an example of DAPPLE's output.

In this case study, we use DAPPLE to illustrate the process of, and provide guidance for, the use of homology for translating knowledge of P-sites across species, in particular across large evolutionary distances. We address several questions applicable to this process, including the following:

- *Question #1: Which P-site database(s) should be used?* The first step in the homology-based identification of P-sites is the selection of one or more reference databases. Multiple databases dedicated to experimentally characterized P-sites exist; however, it is not obvious how their contents relate, or the degree to which they overlap.
- *Question #2: How does the number of sequence differences between a query 15-mer and its best match in the target proteome affect the likelihood that the best match is an actual P-site?* It seems reasonable to believe that the fewer the sequence differences, the greater the likelihood; however, a quantitative relationship has not yet been ascertained.
- *Question #3: From which species should known P-sites be selected?* We hypothesize that using known P-sites from species that are closely related to the target species will improve the ability to predict P-sites.
- *Question #4: Is it possible to identify P-sites in species that are distantly related to those from which known sites are available?* We previously illustrated the use of DAPPLE by predicting bovine P-sites [14]. Bovine represented a relatively 'easy' test case, as it is closely related to most of the species that are well represented in the P-site databases. However, the efficacy of homology-based approaches for identifying P-sites in more distantly related species is unclear.
- *Question #5: What criteria should be used in selecting putative P-sites?* As illustrated in Table 1, DAPPLE outputs a large amount of information for each experimentally determined P-site used as input. The effective use of such data in selecting probable P-sites in the target species is discussed.

For a number of reasons, we chose the honeybee (*Apis mellifera*) as the target species in this study. First, except for a handful of sites in honeybee venom [15], to our knowledge there are no experimentally determined P-sites for *A. mellifera*. Second, the genome of *A. mellifera* has been sequenced [16], which DAPPLE requires. Third, *A. mellifera* is distantly related to most of the species represented in the P-site databases, allowing question #4 to be addressed. Fourth, the honeybee is important from an economic and ecological perspective [17]. The number of honeybee colonies around the world has recently been declining, prompting efforts to discover causes of, and solutions to, this phenomenon [18].

This case study focuses on one particular posttranslational modification (phosphorylation). However, much of the content would also be applicable to other posttranslational modifications, or even to other sequence features involving motifs.

Table 1. Example of a DAPPLE output table

1	2	3	4	5	6	7	8	9	10
Q9Y678	Human	IIVTSSLTKDMTGKE	T110	T110	H9KB37	IIVTSSLTKDMTGKE	0	0	Y
A5GFW1	Pig	WSVHAPSSRRTTLCG	S284	S199	H9K453	WSVHAPSSKRNTLCG	2	2	N
D3YZP9	Mouse	RQLSESESSLEMDDE	S320	S299	H9JYV3	RHLSESESSLEMEEE	3	0	Y
P05131	Bovine	EEEDIRVSITEKCGK	S339	S341	H9KJF2	EEETLRISLTEKCAK	5	4	Y
P00523	Chicken	RLIEDNEYTARQGAK	Y416	Y388	H9K357	RLIKEDEYEARIGA	6	4	Y

Some columns have been omitted because of space constraints. The meanings of the columns are: (1) accession number of the full protein containing the query peptide; (2) organism encoding the protein containing the query peptide; (3) query peptide sequence; (4) location of the phosphorylated residue in the full protein containing the query peptide; (5) location of the phosphorylated residue in the full protein containing the hit peptide; (6) accession number of the full protein containing the hit peptide; (7) hit peptide sequence; (8) number of sequence differences between the query peptide and the hit peptide; (9) number of non-conservative sequence differences between the query peptide and the hit peptide; (10) 'Y' if the proteins represented by the accession numbers in columns 1 and 6 are orthologs according to the reciprocal BLAST hits method, and 'N' otherwise. The first four columns are part of the input to DAPPLE, whereas the remaining columns are calculated by DAPPLE.

Methods

Question #1: Which P-site database(s) should be used?

Data from four P-site databases (PhosphoSitePlus [4, 5], Phospho.ELM [19–21], P^3DB [6, 7] and PhosphoGRID [22, 23]) were downloaded. Each database record contained information about a single P-site from a given species. The proteomes of all species represented in the databases, as well as that of *A. mellifera*, were downloaded from UniProt [24].

To represent all of the known P-site data in a common format, the files from Phospho.ELM, P^3DB and PhosphoGRID were converted into the format used by PhosphoSitePlus. The databases were processed as follows to allow them to be used as input to DAPPLE.

- If a record contained a UniProt accession number, but that accession number was not found in the UniProt proteome for the associated species, then that record was deleted.
- All P-sites were represented by peptides of 15 amino acids in length. Where possible, each peptide was composed of the P-site at its center plus seven residues on either side. If a given P-site was within seven residues of the N-terminus or the C-terminus, the peptide was composed of the first or last 15 residues of the full protein, respectively.
- Any record whose corresponding 15-mer contained an ambiguous amino acid was removed.
- Each record included the accession number of the protein in which the P-site described by that record is found. DAPPLE requires that these accession numbers match those in the corresponding proteomes. As described above, the proteomes downloaded were from UniProt. Thus, DAPPLE requires that the records contain UniProt accession numbers, a requirement already met by PhosphoSitePlus and Phospho.ELM. In contrast, records in PhosphoGRID and P^3DB did not necessarily have UniProt accession numbers. Where possible, the UniProt accession number corresponding to each record in these databases was determined using the methods described in Supplementary File S1. Records for which no corresponding UniProt accession number could be determined were deleted.

Finally, for a given species, there had to be at least one database containing 10 or more records for that species. Records for species not meeting this criterion were deleted. Figure 1 contains a flowchart illustrating the above procedure.

The overlap in the four P-site databases was analyzed from two perspectives: the species represented in each database, and—for each species represented in more than one database—the number of P-sites from that species that were shared or unique. Venn diagrams were created using the R package *VennDiagram* to visualize this information.

Question #2: How does the number of sequence differences between a query 15-mer and its best match in the target proteome affect the likelihood that the best match is an actual P-site?

This question could be answered for a species *S* by running DAPPLE using as input known P-sites from species other than *S*, and using *S* as the target species. For a given number of sequence differences *n*, the percentage of matches that were known to be P-sites in *S* could be calculated. However, this method assumes that all P-sites in *S* have been experimentally determined—otherwise, if the best match *B* for a given known P-site was not a known site in *S*, it is not clear whether this is because *B* is not a P-site, or because *B* is a P-site that has yet to be discovered.

Unfortunately, there are no species for which the complete phosphoproteome has been determined. As such, we performed the above procedure with *S* = human, which is the species having the most well-characterized phosphoproteome. For each non-human species represented in our filtered dataset of P-sites, the known P-sites from that species were searched using DAPPLE against the human proteome. For each value of *n*, the percentage of matches with that value of *n* that were known human P-sites was determined.

For completeness, three variants of the above procedure were performed. The first involved calculating the number of *non-conservative* sequence differences (that is, a mismatch was counted only if the two amino acids had a BLOSUM62 score less than zero). The second variant involved using the mouse proteome as the target rather than the human proteome. The third variant was a combination of the first two (calculating non-conservative substitutions against the mouse proteome).

Question #3: From which species should known P-sites be selected?

DAPPLE was run using the known P-sites that remained after the filtering steps described above, and with the *A. mellifera* proteome as the target. The following values were calculated for each species represented in the P-site databases:

- The percentage of known P-sites with two or fewer sequence differences between the corresponding 15-mer peptide and its best match in the honeybee proteome ('category A matches');
- the percentage with between three and six sequence differences ('category B matches');
- the percentage with seven or more sequence differences ('category C matches'); and
- the percentage for which the full protein corresponding to the query peptide, and the full protein corresponding to the query peptide's best match in the target proteome, were reciprocal BLAST hits.

The first three percentages sum to 100%, while the final percentage is independent of the others. If a 15-mer was found more than once in the same species (either because it was in multiple databases, or because it was found in multiple proteins from that species), then it was counted only once.

We also determined whether the phylogenetic relatedness to honeybee of each species represented in the P-site databases correlated with the level of P-site conservation between that species and honeybee. To estimate phylogenetic relatedness, the mitochondrial genome sequence was downloaded from GenBank for honeybee, as well as for each species represented in the P-site databases (except for the plants and yeast, whose mitochondria are not comparable with those from animals [25]). The EMBOSS program *needle* was used to perform a pair-wise global alignment between the honeybee mitochondrial genome sequence and the mitochondrial sequence from each species represented in the P-site databases. The percent identity for each pair of sequences was ascertained, and we determined the correlation between the percent identity of the mitochondrial genome sequences and the percentage of known P-sites with category A matches in the honeybee proteome.

Question #4: Is it possible to identify P-sites in species that are distantly related to those from which known sites are available?

Here, the same data used to answer question #3 were examined, except rather than analyzing the known P-sites from each

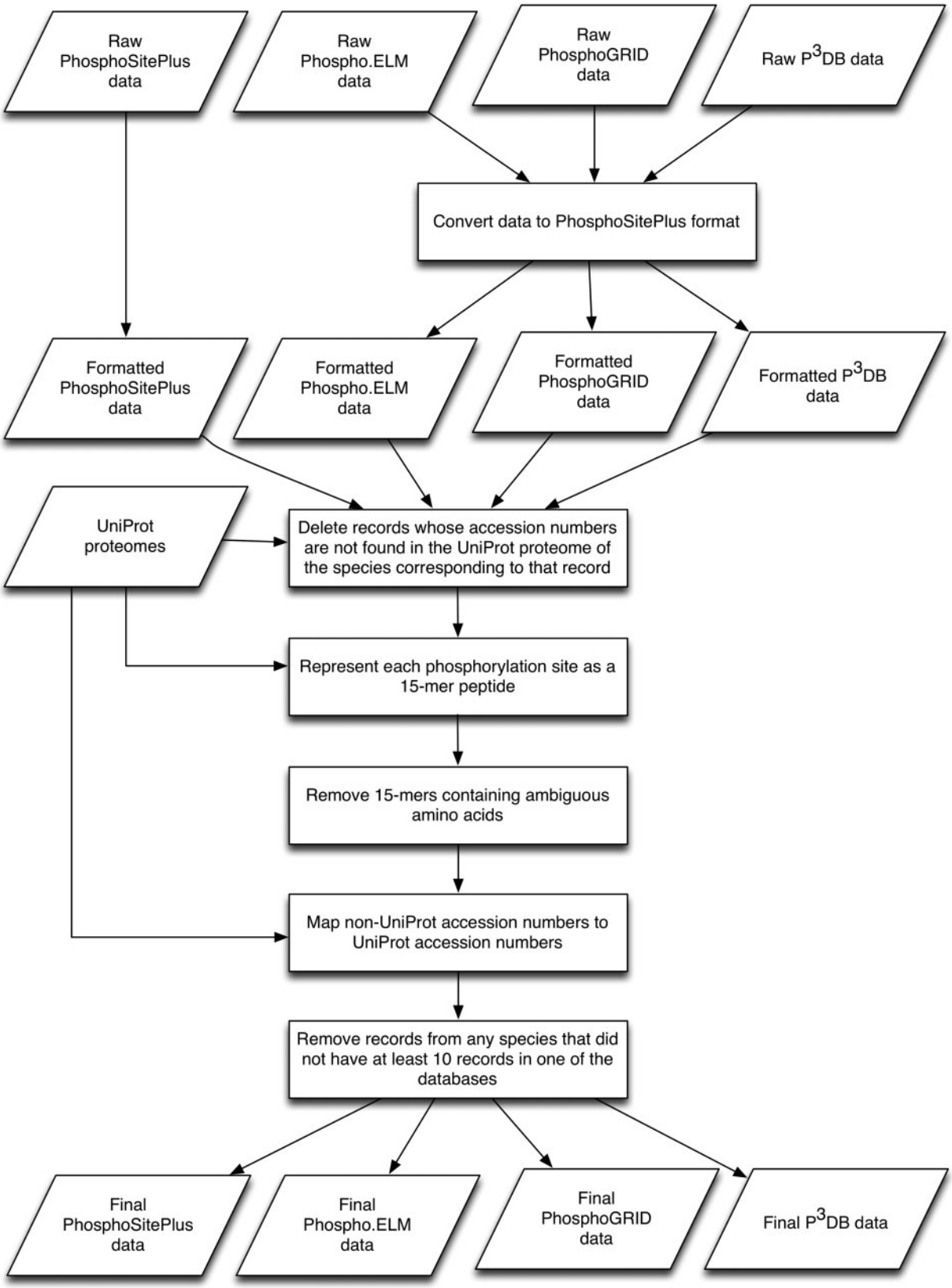

Figure 1. Flowchart depicting the procedure used to process and filter the data in the four P-site databases.

Table 2. Number of P-sites for each species in each P-site database after filtering using the procedures described in Methods

Species	PhosphoSitePlus	Phospho.ELM	P^3DB	PhosphoGRID
Homo sapiens (human)	150 612	35 425	0	0
Mus musculus (mouse)	68 062	7255	0	0
Rattus norvegicus (rat)	9358	544	0	0
Medicago truncatula	0	0	13 515	0
Arabidopsis thaliana	0	0	14 791	0
Oryza sativa (rice)	0	0	7850	0
Saccharomyces cerevisiae (yeast)	0	57	0	6440
Caenorhabditis elegans	0	1470	0	0
Drosophila melanogaster	0	2278	0	0
Glycine max (soybean)	0	1	2092	0
Vitis vinifera (grape)	0	0	850	0
Bos taurus (bovine)	463	188	0	0
Gallus gallus (chicken)	334	102	0	0
Oryctolagus cuniculus (rabbit)	169	89	0	0
Sus scrofa (pig)	80	18	0	0
Zea mays (corn)	0	3	107	0
Xenopus laevi (frog)	33	0	0	0
Canis lupus familiaris (dog)	40	5	0	0
Ovis aries (sheep)	11	12	0	0
Clupea pallasii (pacific herring)	0	10	0	0

species separately, they were considered collectively to characterize the overall effectiveness of using homology to identify putative P-sites in honeybee.

Question #5: What criteria should be used in selecting putative P-sites in the target species?

The output table from DAPPLE contains several columns that help to evaluate the quality of a given match. Based on previous experience in selecting predicted P-sites for constructing a honeybee-specific kinome microarray [26], we make several recommendations for selecting matches that are likely to represent real phosphorylation events. We also provide rationales for the inclusion of three of the peptides on this microarray.

Results and discussion

Question #1: Which P-site database(s) should be used?

Often, two or more databases store biological data of a given type. For instance, general sequence data are present in databases maintained by both the National Center for Biotechnology Information (NCBI) and the European Bioinformatics Institute (EBI), and there are multiple databases containing 16S rRNA sequences [27–29]. Although there can be benefits associated with multiple databases, it can also produce confusion (for example, it may not be clear how the data in the different databases relate) and frustration (for example, different databases usually store their data in different formats and with different identifiers, making it difficult to combine their data).

Given that there are multiple P-site databases, a user may wonder how their contents compare. Here, four major P-site databases (PhosphoSitePlus [4, 5], Phospho.ELM [19–21], P^3DB [6, 7] and PhosphoGRID [22, 23]) were compared both at the species level (that is, the number of P-sites from a given species in each database) and at the sequence level (if multiple databases had sites from the same species, to what degree did those sites overlap?).

Table 2 shows that the level of overlap at the species level among the four databases was small. For instance, after filtering, PhosphoGRID contained 6440 P-sites from *Saccharomyces cerevisiae* (the only species represented in this database), while Phospho.ELM—the only other database containing records from *S. cerevisiae*—contained just 57. With two minor exceptions (one soybean site and three corn sites in Phospho.ELM), the species represented in P^3DB were not represented elsewhere. The only two databases that had significant overlap were PhosphoSitePlus and Phospho.ELM. Of the 16 species represented in at least one of these databases, nine were represented in both, six were represented only in Phospho.ELM and one was represented only in PhosphoSitePlus. For most of the species represented in both databases, PhosphoSitePlus contained many more sites than Phospho.ELM. For instance, PhosphoSitePlus contained >150 000 human sites versus <36 000 in Phospho.ELM. The ratios of mouse and rat sites were even more biased in favor of PhosphoSitePlus: 68 062 versus 7255 and 9358 versus 544, respectively. Of the six species that were present in Phospho.ELM but not PhosphoSitePlus, four had only a few sites; however, Phospho.ELM contained 2278 sites from *D. melanogaster* and 1470 from *Caenorhabditis elegans*—species absent from the other three databases. These data suggest that none of the four databases is rendered completely redundant by any of the others.

Given that PhosphoSitePlus and Phospho.ELM contained sites from many of the same species, we determined the degree to which the P-sites from a given species in PhosphoSitePlus overlapped with those in Phospho.ELM. For each of the nine species represented in both databases, the number of sites found only in PhosphoSitePlus, only in Phospho.ELM or in both databases was determined. Venn diagrams representing these results are given in Figure 2, showing that, for most species, the majority of the sites in Phospho.ELM were also in PhosphoSitePlus. However, Phospho.ELM had some unique sites in eight of the nine species.

Given the information in Table 2 and Figure 2, it appears that the most appropriate database to use for the homology-based

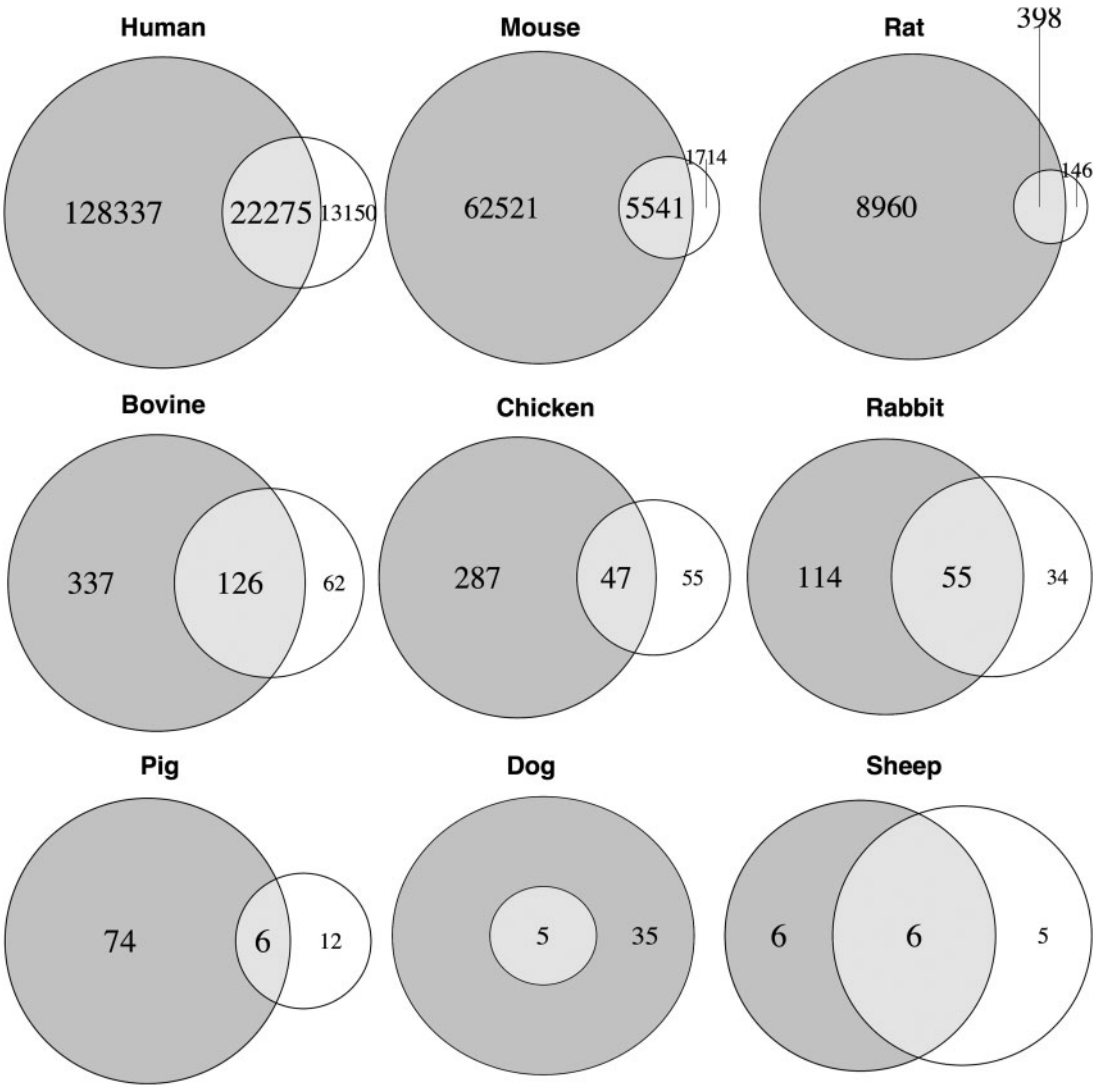

Figure 2. The number of P-sites found in PhosphoSitePlus only (dark gray), Phospho.ELM only (white) or both databases (light gray) for each of the nine species that were represented in both databases. The sets are drawn to scale within each species, but not between species.

prediction of P-sites depends on the target species. If that species is a plant, then P^3DB would be most appropriate; if it is closely related to yeast, then PhosphoGRID should be chosen. If the target species is mammalian, PhosphoSitePlus would be preferable over Phospho.ELM because of its greater number of sites; however, there is nothing precluding the use of data from both. If the species of interest is closely related to *C. elegans* or *D. melanogaster*, then Phospho.ELM ought to be used, as it is the only database containing sites from these species. Finally, there is no apparent disadvantage to using all four databases. In fact, the data presented in subsequent sections suggest that one should not limit the selection of known P-sites to those from species that are ostensibly 'closely related' to the target species.

Question #2: How does the number of sequence differences between a query 15-mer and its best match in the target proteome affect the likelihood that the best match is an actual P-site?

DAPPLE was used to search known nonhuman P-sites against the human proteome, and the percentage of matches with a given number of sequence differences that are known to be human P-sites was determined. The results are given in Table 3, which suggests that the likelihood of a particular match being a P-site declines steadily with the number of sequence differences associated with that match. For instance, when known rat P-sites were searched against the human proteome, ~60% of the matches with zero sequence differences were known human P-sites, versus 46% for three sequence differences and 31% for six. This trend was relatively consistent across species, although the percentages were generally lower for species more distantly related to human. For example, when known *A. thaliana* P-sites were searched against the human proteome, just 14% of the matches with six sequence differences were known human sites, compared with 31% for rat.

Similar trends were observed when mouse was used as the target species rather than human (Supplementary Table S1), and when non-conservative sequence differences were calculated rather than sequence differences (Supplementary Tables S2 and S3). Supplementary Tables S4–S7 contain the number of peptides having a given number of sequence

Table 3. Relationship between the number of sequence differences between a known nonhuman P-site and that P-site's best match in the human proteome, and the likelihood that the best match is itself a known P-site

Species	Number of sequence differences										
	0	1	2	3	4	5	6	7	8	9	10
Mus musculus (mouse)	53.4	48.9	43.6	39.2	34.7	29.2	23.0	17.9	12.5	5.7	0.0
Rattus norvegicus (rat)	59.7	56.4	54.1	46.2	39.6	32.0	31.3	20.1	12.0	4.1	0.0
Medicago truncatula	58.5	61.5	44.3	39.4	32.3	22.6	17.3	16.4	12.9	5.8	0.0
Arabidopsis thaliana	75.0	70.0	65.7	45.6	26.7	15.4	13.9	13.9	10.4	5.0	1.3
Oryza sativa (rice)	0.0	58.8	57.7	48.1	36.7	23.7	16.7	13.9	12.5	5.7	0.0
Saccharomyces cerevisiae (yeast)	37.5	50.0	40.5	40.0	34.8	21.1	17.3	15.5	13.5	6.3	4.0
Caenorhabditis elegans	58.3	68.4	57.9	44.0	20.8	22.7	22.0	19.1	12.2	4.4	0.0
Drosophila melanogaster	93.3	72.2	78.6	48.1	27.6	25.4	24.7	20.1	16.7	3.8	0.0
Glycine max (soybean)	100.0	100.0	57.1	42.9	61.1	11.5	20.9	22.2	12.7	6.7	0.0
Vitis vinifera (grape)	66.7	50.0	71.4	60.0	60.0	21.1	20.0	17.2	13.4	6.2	0.0
Bos taurus (bovine)	68.9	57.0	37.3	52.8	20.0	21.4	32.4	18.2	18.8	0.0	–
Gallus gallus (chicken)	71.7	74.4	66.7	36.1	58.6	48.0	28.6	23.1	24.1	20.0	–
Oryctolagus cuniculus (rabbit)	76.0	64.7	50.0	38.5	25.0	66.7	33.3	0.0	50.0	–	–
Sus scrofa (pig)	78.9	80.0	62.5	0.0	40.0	14.3	0.0	33.3	0.0	16.7	–
Zea mays (corn)	–	–	–	100.0	100.0	40.0	30.8	17.6	20.6	0.0	–
Xenopus laevi (frog)	100.0	–	100.0	66.7	66.7	66.7	0.0	10.0	0.0	0.0	–
Canis lupus familiaris (dog)	72.0	83.3	50.0	0.0	0.0	0.0	–	–	–	–	–
Ovis aries (sheep)	75.0	60.0	0.0	0.0	100.0	100.0	–	–	–	–	–
Clupea pallasii (pacific herring)	–	–	–	–	66.7	100.0	100.0	0.0	–	–	–

Known P-sites (represented as 15-mer peptides with the phosphorylated residue in the center) from the species listed in the first column were searched against the human proteome using DAPPLE, and the number of sequence differences between a known P-site and its best match in the human proteome was recorded. The remaining columns indicate the percentage of matches with a given number of sequence differences that are known to be human P-sites. If a cell contains '–', it means that there were no peptides from that species whose best match had the indicated number of sequences differences in the human proteome. If a cell is shaded, it indicates that the number of peptides having that number of sequence differences (irrespective of the percentage that are known to be human P-sites) was ≥100, meaning that greater confidence can be placed in the values in these cells. Unshaded cells may contain anomalous values owing to small sample sizes. For example, there were only two peptides from soybean having matches in the human proteome with zero sequence differences, both of which were known human P-sites, giving an anomalous value of 100%. In contrast, the value for mouse is far more representative: 13 710 of the 25 663 peptides with zero sequence differences (53.4%) in the human proteome were known human P-sites. Supplementary Table S4 contains the exact number of peptides that had a given number of sequence differences for each species.

differences (or non-conservative sequence differences) in the human or mouse proteome.

Although a large number of P-sites are known in human, there are undoubtedly many that have yet to be discovered. Thus, the percentages in Table 3 must be interpreted carefully. For example, it is unlikely that only 60% of human matches to rat P-sites with zero sequence differences are real P-sites—the true number is certainly higher. However, Table 3 does provide insight into the *relative* likelihood of a particular match being a true P-site. For example, based on the rat data, it appears that the likelihood of a match with six sequence differences being a real P-site is approximately half the likelihood of a match with zero sequence differences being a real P-site.

The results presented in Table 3 and Supplementary Tables S1–S7 suggest that a large proportion of the P-sites predicted by DAPPLE—particularly those with few sequences differences—are likely to be real P-sites, suggesting that this homology-based strategy is useful for identifying biologically likely P-sites. As a complement to this validation, it would also be of interest to perform a direct comparison to machine learning-based predictors. Unfortunately, such a comparison would be quite difficult, as traditional measures of accuracy like sensitivity and specificity are problematic for a homology-based approach. A more detailed discussion of these difficulties is available in Supplementary File S2. Despite these difficulties, it may be possible to compare DAPPLE with machine learning-based tools using nontraditional metrics of accuracy and usefulness. This is the subject of ongoing future work.

Question #3: From which species should known P-sites be selected?

DAPPLE was used to identify potential P-sites in the honeybee proteome for each P-site listed in Table 2. Table 4 summarizes the percentage of P-sites from a given species whose corresponding 15-mer had a given number of sequence differences with its best match in the honeybee proteome. Supplementary Table S8 contains the same data, except for non-conservative sequence differences, and Supplementary File S3 contains the complete set of DAPPLE predictions. For all species, the percentage of their known P-sites having category A matches in the honeybee proteome was small. For instance, just 3.9% of human sites had a match with two or fewer sequence differences. These percentages were similar for most other mammals. Unsurprisingly, the plants had lower percentages of category A matches compared with the animals; for instance, just 0.4% of sites from *A. thaliana* had category A matches. Unexpectedly, *D. melanogaster*—the most closely related species to honeybee of the species in Table 2—was no better than the mammals in terms of P-site conservation with honeybee.

To determine whether the evolutionary relatedness between a given species and honeybee correlated with the level of P-site conservation between them, the mitochondrial genome sequence was downloaded for each species, and the percent identity between a given species' mitochondrial genome sequence and that of the honeybee was determined. These values are given in the last column of Table 4. The correlation between the percentage of P-sites in a given species having a category A match in the honeybee proteome (the second column in

Table 4. Degree of P-site conservation between *A. mellifera* and each species represented in the P-site databases

Species	Sequence differences			RBH? (%)	mt % identity
	0–2 (%)	3–6 (%)	7+ (%)		
Homo sapiens (human)	3.9	16.4	79.7	9.4	32.8
Mus musculus (mouse)	3.0	14.5	82.5	7.5	37.0
Rattus norvegicus (rat)	5.5	16.7	77.9	11.1	37.1
Medicago truncatula	1.1	10.3	88.7	1.8	N/A
Arabidopsis thaliana	0.4	10.0	89.5	0.9	N/A
Oryza sativa (rice)	0.8	9.2	90.1	0.9	N/A
Saccharomyces cerevisiae (yeast)	1.1	11.5	87.4	3.3	N/A
Caenorhabditis elegans	2.9	15.5	81.6	7.4	47.2
Drosophila melanogaster	5.0	18.4	76.6	13.1	52.8
Glycine max (soybean)	0.9	10.9	88.2	1.8	N/A
Vitis vinifera (grape)	0.5	11.8	87.7	1.9	N/A
Bos taurus (bovine)	4.6	13.0	82.4	8.8	35.0
Gallus gallus (chicken)	5.2	18.9	75.9	13.6	30.4
Oryctolagus cuniculus (rabbit)	7.2	13.8	79.0	12.2	37.7
Sus scrofa (pig)	12.8	12.8	74.4	8.1	32.7
Zea mays (corn)	0.0	9.8	90.2	3.9	N/A
Xenopus laevi (frog)	6.1	3.0	90.9	0.0	N/A
Canis lupus familiaris (dog)	0.0	12.8	87.2	2.6	37.0
Ovis aries (sheep)	11.8	23.5	64.7	0.0	37.1
Clupea pallasii (pacific herring)	0.0	100.0	0.0	0.0	32.8

The second, third and fourth columns list the percentage of known P-sites from the species in the first column that had 0–2 sequence differences, 3–6 differences or 7 or more differences, respectively, between a given 15-mer sequence and its best match in the honeybee proteome. The 'RBH?' column lists the percentage of sites predicted by DAPPLE to be reciprocal BLAST hits [see 14]. The final column lists the percent identity between the sequence of the mitochondrial (mt) genome from that species and the sequence of the honeybee mitochondrial genome. Values were not determined for cells marked 'N/A'; this was either because the species' mitochondrial genome was not comparable with that of honeybee (for plants and yeast) or because the complete mitochondrial genome sequence was not available (for *Xenopus laevi*).

Table 4) and the percent identity of the mitochondrial genome sequences was determined. Interestingly, the correlation was near zero ($r = -0.10$). This lack of correlation is unlikely to be the case in general; its presence here is likely because all of the species represented in the P-site databases are distantly related to the target species (honeybee).

In identifying P-sites in the honeybee proteome, a user might be tempted to use only known P-sites from *D. melanogaster*, which is the most closely related species to honeybee among those represented in the P-site databases. However, there are at least two reasons why this may be unwise. First, *D. melanogaster* is not closely related to honeybee in an absolute sense—although they belong to the same class (Insecta) in the taxonomic hierarchy, they belong to different orders (Diptera and Hymenoptera, respectively); further, their mitochondrial DNA sequences are only 53% identical (Table 4). Second, Table 4 suggests that the level of P-site conservation between *D. melanogaster* and honeybee is no better than between the mammals and honeybee. Therefore, if a researcher used only known sites from *D. melanogaster*, few potential honeybee sites would be discovered. Specifically, using only known P-sites from *D. melanogaster* would give ~100 category A matches; in comparison, nearly 6000 category A matches were discovered using known human sites. Depending on the research problem at hand, 100 matches may be sufficient; conversely, if specific signaling pathways are of interest, this would provide little choice in selecting peptides.

In answering question #1, we did not explicitly compare the results obtained when using each individual database as input to DAPPLE. However, by considering the results of the database comparison (question #1) and the species comparison (question #3) together, it becomes clear how the results would differ depending on the database used. For instance, if P^3DB (which contains only plant P-sites) were used on its own, then a small percentage of the known P-sites would have category A matches in the honeybee proteome, as indicated by Table 4 (for example, see the rows for *A. thaliana* and *Oryza sativa*). In contrast, the use of PhosphoSitePlus would result in a larger number of putative honeybee sites, given that this database contains a huge number of human sites (Table 2) and that a similar percentage of human sites had category A matches as compared with other species (Table 4).

Question #4: Is it possible to identify P-sites in species that are distantly related to those from which known sites are available?

When all of the P-site data for the different species were combined, the percentage of known sites that had category A matches (two or fewer sequence differences) in the honeybee proteome was <3%. This differs markedly from the bovine proteome, for which ~60% of sites had a match with two or fewer sequence differences [14]. This likely reflects the fact that bovine is much more closely related to the species represented in the P-site databases than honeybee.

Despite the low percentage of category A matches, the sheer number of sites contained in the P-site databases means that the homology-based approach was still able to identify many putative honeybee sites. In total, nearly 9000 known P-sites had category A matches in the honeybee proteome. Table 3 suggests that these matches have a good likelihood of being real P-sites. Thus, in addition to being able to identify P-sites in species closely related to those represented in the P-site databases, DAPPLE also appears to be useful for identifying sites in more distantly related species.

Question #5: What criteria should be used in selecting putative P-sites in the target species?

DAPPLE outputs many pieces of information about each known P-site and its best match in the target proteome. The most important piece is likely the number of sequence differences between the query peptide and its best match in the target proteome (column 8 in Table 1), the details of which were investigated in Question #2. Two additional criteria that appear relevant to selecting biologically likely P-sites are as follows.

- The locations of the query site and the hit site in their respective full proteins (columns 4 and 5 in Table 1). It seems more likely that a match would represent a real P-site if the two locations are close together (e.g. Y352 and Y358) than if they are far apart (e.g. Y352 and Y15).
- Whether the full proteins are reciprocal BLAST hits (column 10 in Table 1). This information is useful because, if the two proteins are orthologs, then it is more likely that a conservation of function exists between the two proteins, in turn making it more likely that functional elements (such as P-sites) would also be conserved.

One application for the prediction of P-sites is the design of species-specific kinome microarrays. Each spot on a kinome microarray contains a population of peptides of a particular

sequence, where that sequence contains (usually as its central residue) a site that is known or suspected to be phosphorylated. Recently, we used DAPPLE to predict P-sites in the honeybee using experimentally determined sites from other organisms for the purposes of designing a honeybee-specific kinome array [26]. From these sites, the criteria described above were used to select 299 peptides for inclusion on the array. The following three examples illustrate how the information provided by DAPPLE was used to select peptides.

- DLDHERMSYLLYQML—The central residue in this peptide corresponds to residue S135 in the honeybee protein with UniProt accession number H9KH67. The known P-site that was used to identify this peptide was S129 in the mouse protein with accession number Q91Y86, which corresponds to the peptide ELDHERMSYLLYQML. There were a number of reasons why DLDHERMSYLLYQML was chosen. First, there was only one sequence difference between this peptide and its mouse counterpart. Second, the proteins with accession numbers H9KH67 and Q91Y86 were reciprocal BLAST hits. Third, the location of the P-site in the mouse protein (residue 129) was close to its location in the honeybee protein (residue 135), giving further evidence of the correspondence between the two sites.
- HKLGGGQYGDVYEAV—The central residue corresponds to residue Y300 in the honeybee protein H9K2C5. The P-site Y253 in the human protein with accession number P00519 (which corresponds to the 15-mer HKLGGGQYGEVYEGV) was used to identify it. This peptide was chosen because the residue locations were reasonably similar (300 versus 253), the number of sequence differences between the two peptides was small (two) and the corresponding full proteins are reciprocal BLAST hits.
- YKERIDEYDYAKPLE—The central residue of this peptide corresponds to residue Y1510 in the honeybee protein H9KFK6. It was identified via the known P-site Y596 in the rat protein with accession number Q920L2, which corresponds to the 15-mer YKVRIDEYDYSKPIE. Compared with the above peptides, this choice was more speculative, as the phosphorylated residues were far apart, the proteins were not reciprocal BLAST hits, and the similarity between the honeybee peptide and the rat peptide was lower (three sequence differences).

The peptides chosen for our honeybee-specific peptide array did appear to contain real P-sites, as meaningful, characteristic phosphorylation patterns were detected when the arrays were exposed to bees of different developmental stages and/or phenotypes [26]. Using these arrays, we successfully elucidated signaling mechanisms associated with resistance or susceptibility to *Varroa destructor*, a mite thought to be a major contributor to the collapse of honeybee colonies.

Conclusion

In this case study, we provided practical guidance for the homology-based prediction of P-sites in a target species (in this case, honeybee) using known P-sites from other species. Four major P-site databases were compared, and it was found that their contents differed substantially, justifying the use of data from all four. To validate the use of DAPPLE in identifying P-sites, we showed that known P-sites from many species can be used to successfully identify human P-sites, and that the number of sequence differences is related to the likelihood of the match being a real P-site. Although only a small percentage of known P-sites had good matches in the honeybee proteome, the sheer number of known P-sites meant that thousands of honeybee P-sites were identified, making this homology-based approach useful even for species that are distantly related to those represented in the P-site databases. The percentage of known sites with good matches in the honeybee proteome was relatively consistent among the different organisms. This was true even for *D. melanogaster*, which users might expect to have greater P-site conservation with honeybee. This suggests that, when predicting P-sites in distantly related species, data from all organisms in the P-site databases should be used, not just from those that are most closely related to the target species.

Supplementary data

Supplementary data are available online at http://bib.oxfordjournals.org/.

Key Points

- While the P-site databases tested have some overlapping content, it is worthwhile to use data from all four when performing homology-based prediction of P-sites.
- The likelihood that a particular DAPPLE match is a real P-site appears to be closely related to the number of sequence differences between that match and the query peptide.
- The (perceived) level of evolutionary relatedness between the target species and a species from which known P-sites are available is not necessarily a good indication of the degree of P-site conservation between them.
- Owing to the large number of known P-sites, it is possible to make useful homology-based predictions even in evolutionarily distant species.

Funding

Funding was provided by Genome Canada and the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

1. Hunter T, Plowman GD. The protein kinases of budding yeast: six score and more. *Trends Biochem Sci* 1997;**22**: 18–22.
2. Manning G, Whyte DB, Martinez R, *et al*. The protein kinase complement of the human genome. *Science* 2002;**298**:1912–34.
3. Caenepeel S, Charydczak G, Sudarsanam S, *et al*. The mouse kinome: discovery and comparative genomics of all mouse protein kinases. *Proc Natl Acad Sci USA* 2004;**101**:11707–12.
4. Hornbeck PV, Chabra I, Kornhauser JM, *et al*. PhosphoSite: a bioinformatics resource dedicated to physiological protein phosphorylation. *Proteomics* 2004;**4**:1551–61.
5. Hornbeck PV, Kornhauser JM, Tkachev S, *et al*. PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res* 2012;**40**:D261–70.
6. Gao J, Agrawal GK, Thelen JJ, *et al*. P3DB: a plant protein phosphorylation database. *Nucleic Acids Res* 2009;**37**:D960–2.

7. Yao Q, Bollinger C, Gao J, *et al*. P(3)DB: an integrated database for plant protein phosphorylation. *Front Plant Sci* 2012;**3**:206.
8. Xue Y, Gao X, Cao J, *et al*. A summary of computational resources for protein phosphorylation. *Curr Protein Pept Sci* 2010;**11**:485–96.
9. Trost B, Kusalik A. Computational prediction of eukaryotic phosphorylation sites. *Bioinformatics* 2011;**27**:2927–35.
10. Jalal S, Arsenault R, Potter AA, *et al*. Genome to kinome: species-specific peptide arrays for kinome analysis. *Sci Signal* 2009;**2**:pl1.
11. Yaffe MB, Leparc GG, Lai J, *et al*. A motif-based profile scanning approach for genome-wide prediction of signaling pathways. *Nat Biotechnol* 2001;**19**:348–53.
12. Xue Y, Ren J, Gao X, *et al*. GPS 2.0, a tool to predict kinase-specific phosphorylation sites in hierarchy. *Mol Cell Proteomics* 2008;**7**:1598–608.
13. Biswas AK, Noman N, Sikder AR. Machine learning approach to predict protein phosphorylation sites by incorporating evolutionary information. *BMC Bioinformatics* 2010;**11**:273.
14. Trost B, Arsenault R, Griebel P, *et al*. DAPPLE: a pipeline for the homology-based prediction of phosphorylation sites. *Bioinformatics* 2013;**29**:1693–5.
15. Li R, Zhang L, Fang Y, *et al*. Proteome and phosphoproteome analysis of honeybee (*Apis mellifera*) venom collected from electrical stimulation and manual extraction of the venom gland. *BMC Genomics* 2013;**14**:766.
16. Honeybee Genome Sequencing Consortium. Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature* 2006;**443**:931–49.
17. Calderone NW. Insect pollinated crops, insect pollinators and US agriculture: trend analysis of aggregate data for the period 1992-2009. *PLoS One* 2012;**7**:e37235.
18. Dietemann V, Nazzi F, Martin SJ, *et al*. Standard methods for varroa research. *J Apic Res* 2013;**52**:1–54.
19. Diella F, Cameron S, Gemünd C, *et al*. Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins. *BMC Bioinformatics* 2004;**5**:79.
20. Diella F, Gould CM, Chica C, *et al*. Phospho.ELM: a database of phosphorylation sites–update 2008. *Nucleic Acids Res* 2008;**36**:D240–4.
21. Dinkel H, Chica C, Via A, *et al*. Phospho.ELM: a database of phosphorylation sites–update 2011. *Nucleic Acids Res* 2011;**39**:D261–7.
22. Stark C, Su TC, Breitkreutz A, *et al*. PhosphoGRID: a database of experimentally verified in vivo protein phosphorylation sites from the budding yeast *Saccharomyces cerevisiae*. *Database* 2010;**2010**:bap026.
23. Sadowski I, Breitkreutz BJ, Stark C, *et al*. The PhosphoGRID *Saccharomyces cerevisiae* protein phosphorylation site database: version 2.0 update. *Database* 2013;**2013**:bat026.
24. UniProt Consortium. Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Res* 2013;**41**:D43–7.
25. Christensen AC. Plant mitochondrial genome evolution can be explained by DNA repair mechanisms. *Genome Biol Evol* 2013;**5**:1079–86.
26. Robertson AJ, Trost B, Scruten E, *et al*. Identification of developmentally-specific kinotypes and mechanisms of Varroa mite resistance through whole-organism, kinome analysis of honeybee. *Front Genet* 2014;**5**:139.
27. Cole JR, Wang Q, Fish JA, *et al*. Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res* 2014;**42**:D633–42.
28. Quast C, Pruesse E, Yilmaz P, *et al*. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* 2013;**41**:D590–6.
29. DeSantis TZ, Hugenholtz P, Larsen N, *et al*. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Env Microbiol* 2006;**72**:5069–72.