Electronic Supplementary Material (ESI) for Molecular BioSystems

Can Simple Codon Pair Usage Predict Protein-Protein Interaction?

Yuan Zhou, Ying-Si Zhou, Fei He, Jiangning Song and Ziding Zhang

Electronic Supplemental Information

Supplemental Methods

Statistical analyses of codon/codon pair frequency differences between interacting protein pairs

We compared the differences of codon/codon pair frequency between 4,156 interacting protein pairs in the DIP yeast core dataset[1] and the randomly selected negative protein pairs. The latter are 19 times larger than the former. DIP positives were downloaded from the DIP database (http://dip.doe-mbi.ucla.edu/dip/Download.cgi). We searched for informative codon pairs whose frequency differences in the positives were significantly different from those in the negatives [either larger or smaller, according to Welch's t-test followed by Benjamini-Hochberg correction with a certain significance level cutoff (i.e., correted p-value = 0.05)]. In other words, a corrected p-value was calculated for each codon pair to describe if it was similarly used or dissimliarly used in PPIs, therefore informative in discriminating interacting protein pairs and non-interacting protein pairs. We searched informative codons in the same way.

A codon pair could be informative (preferably used in a non-random fashion among interacting protein pairs), either dependent or independent of non-random codon usage and non-random amino acid pair usage. We attempted to estimate the number of independent ones using permutated sequence sets. In a permutated sequence set, only synonymous codons within each coding sequence were shuffled. That is to say, we swapped each codon with one randomly selected synonymous codon (if existed) throughout the coding sequence. Therefore, the distribution of codon pair frequency between interacting protein pairs was altered after such shuffling, while that of codon frequency and amino acid pair frequency remained unchanged. We generated 1,000 permutated sets and collected codon pairs meeting either of the folowing two criteria in each dataset: 1) A codon pair was shown to be similarly used in interacting protein pairs by the test described in the above paragraph, and the permutation would result in a less similar codon pair usage (paired Student's t-test followed by Benjamini-Hochberg correction, $p<0.05$); 2) A codon pair was shown to be dissimilarly used in interacting protein pairs, and permutation would result in a more similar codon pair usage in interacting protein pairs (paired Student's t-test followed by Benjamini-Hochberg correction, $p<0.05$). If a codon pair met one of the two criteria across at least 950 out of 1,000 permutated sets, this codon pair was treated as an informative one that was independent of both non-random codon usage and non-random amino

acid pair usage.

Probing genomic factors that contribute to CCPPI's performance

We compared true positive predictions of different encoding schemes in terms of multiple factors that measure various aspects of interacting protein pairs. The first factor was transcriptional co-expression, which was measured by the fraction of transcriptional co-expressed proteins during cell cycle.[2] Transcriptome data were downloaded from the GEO database (http://www.ncbi.nlm.nih.gov/geo/, accession number GSE4987). The mRNA expression profiles of 781 proteins from the BIOGRID dataset were believed to have significant fluctuations during cell cycle, i.e., they are listed among the top 1,000 periodically expressed genes (http://labs.fhcrc.org/breeden/cellcycle/). 1,258 PPIs among them were shown to be transcriptional co-expressed (i.e., the absolute Pearson's correlation coefficient is larger than 0.5). Similarly, there were 5,489 PPIs among 1,422 proteins co-expressed at the proteome level during yeast cell cycle, according to the protein expression profile[3] retrieved from the PeptideAtlas database (http://www.peptideatlas.org/repository/publications/flory2005/). Finally, functional or subcellular localization similarities between interacting protein pairs were quantified by the RSS values[4] described in the Materials and Methods section from main text and the original paper by Wu *et al.*[4]

Comparison of CCPPI and other methods in the fly dataset

We compared CCPPI with other types of methods, including interolog, domain-domain interaction and phylogenetic profile, in the fruit fly (*Drosophila melanogaster*) dataset. Fruit fly sequences were downloaded from the Ensembl database[7] (http://nov2010.archive.ensembl.org/, version 60). In the case of alternatively spliced genes, the longest coding sequences were used. The known physical interactions in fruit fly and the positive predictions of the interolog method[8] were downloaded from the Interolog Finder database (http://interologfinder.org/), while 150,000 randomly selected non-interacting protein pairs were used as negative benchmarks. The interolog method transferred PPIs from other species to the fruit fly orthologous protein pairs. PPIs were also predicted from 6,074 known domain-domain interactions curated in the iPfam database.[9] More exactly, two fruit fly proteins were predicted to interact if each of them contained one of

the two known interacting domains, respectively. The Pfam domain annotation of each protein was downloaded from the Ensembl database.[7]

Finally, we predicted the fruit fly PPIs using the phylogenetic profile strategy proposed by Sun *et al*[10] with modifications. This method predicted two proteins as interacting partners when they co-occurred/co-disappeared across a set of the reference genomes. We sampled a reference genome by randomly selecting one genome from each clade at certain level of the phylogenetic trees available at NCBI Taxonomy (http://www.ncbi.nlm.nih.gov/Taxonomy/). We noted that there was a prominent unbalance between the total number of sequenced eukaryote genomes and that of sequenced non-eukaryote genomes. Therefore, level 6, level 5 and level 2 were applied to the reference genome sampling from eukaryotes, archaea and bacteria, respectively. As a result, 30 eukaryotes, 65 archaea and 39 bacteria genomes were selected as references (see Table S7 for the full list) and their RefSeq proteins were downloaded from NCBI (http://www.ncbi.nlm.nih.gov/RefSeq/). Then we generated actual/shuffled phylogenetic profile, and calculated mutual information between each protein pairs' phylogenetic profiles according to Sun *et al*.[10] A threshold of 0.15 was used to identify PPIs because none of shuffled profile pair could achieve such a high mutual information level (Fig. S7). We also removed PPIs between the paralogous proteins from fruit fly as described by Sun *et al*.[10]

Supplemental Tables

Table S1 Summary of related and previously published sequence-based PPI prediction methods.

Encoding scheme	Algorithm framework	Main benchmarking dataset	Performance	Ref.
CT encoding	SVM with S kernel	HPRD human positives + the equal number of human homogeneous negatives [a]	Accuracy=83.9%	Ref.[5]
AC encoding	SVM with RBF kernel	DIP yeast positives + the equal number of negatives with different subcellular localization	Accuracy=86.2%	Ref.[6]
Codon frequency difference	Naive Bayes	MIPS yeast positives + about 1.4 million negatives with different subcellular localization	AUC=0.845 [b]	Ref.[11]

[a] The homogeneous negatives were randomly rewired pairs of positive proteins.

[b] This AUC value was estimated from the ROC curve in the original paper.

Table S2 The optimized SVM training parameters.

Encoding scheme	*c* (cost parameter)	*g* (gamma parameter)
CCPPI	64	0.015625
CT encoding	128	0.0625
AC encoding	256	0.0625

We optimized the SVM parameters by 10-fold cross-validation tests using a "DIP+Random" dataset. In particular, a heuristic strategy for searching the optimized parameters was employed. Namely, we started with $c=2^0$ and $g=2^{-10}$, and alternately increased or decreased the value of one parameter by two folds. If a better accuracy was achieved, we then continued to increase or decrease the values of the parameters until no further improvement was observed**.**

Table S3 Performance of CCPPI and the other two encoding schemes when evaluated on the "DIP+RSS Negative" datasets.

Encoding scheme	Accuracy (%)	Precision (%)	Sensitivity (%)	MCC
CCPPI	90.2 ± 0.2	88.5 ± 0.3	92.4 ± 0.3	0.804 ± 0.002
CT encoding	81.3 ± 0.4	81.9 ± 0.5	80.4 ± 0.5	0.627 ± 0.008
AC encoding	73.5 ± 0.3	71.5 ± 0.4	78.2 ± 0.3	0.473 ± 0.006

The datasets comprise of interacting and randomly selected non-interacting protein pairs without any known similar functions or subcellular localizations. The 10-fold cross-validation tests were repeated five times by selecting different negative samples. The results are expressed as mean ± standard deviation. The predictors were trained with the preliminarily optimized parameters.

Table S4 Performance of CCPPI and the other two encoding schemes on the "DIP+Homogenous" datasets.

Encoding scheme	Accuracy (%)	Precision (%)	Sensitivity (%)	MCC
CCPPI	63.7 ± 0.5	66.2 ± 0.7	56.2 ± 1.4	0.278 ± 0.009
CT encoding	59.3 ± 0.4	59.9 ± 0.5	56.5 ± 0.8	0.186 ± 0.007
AC encoding	57.1 ± 0.2	57.6 ± 0.3	53.7 ± 1.4	0.142 ± 0.004

The datasets comprise of interacting and randomly selected non-interacting protein pairs through rewiring of interacting protein pairs. The 10-fold cross-validation tests were repeated five times by selecting different negative samples. The results are expressed as mean ± standard deviation. The predictors were trained with the preliminarily optimized parameters.

Table S5 Performance of different methods/encodings in predicting fruit fly (*Drosophila melanogaster*) PPIs.

Method/Encoding	Cutoff	Sensitivity (%)	Specificity (%)
CCPPI	0.39	15.6%	92.2%
CT encoding	0.68	17.7%	90.7%
AC encoding	0.68	33.0%	81.7%
Interolog	-	3.7%	99.9%
Domain-domain interaction	-	4.3%	99.5%
Phylogenetic profile	0.15	16.8%	91.1%

The dataset was composed of 26,545 interacting protein pairs and 150,000 randomly selected non-interacting protein pairs from the fruit fly genome. See Supplemental Methods for details. The cutoff values corresponding to the 90% specificity in the yeast dataset were used for CCPPI, the CT encoding and the AC encoding.

Table S6 The list of reference genomes used in the phylogenetic profile method.

Eukaryotes	Archaea	Archaea (continue)	Bacteria	Bacteria (continue)
Amphimedon queenslandica	*Acidilobus saccharovorans* 345-15	*Methanohalobium evestigatum* Z-7303	*Acaryochloris marina* MBIC11017	*Nitrospira defluvii*
Apis mellifera	*Aeropyrum pernix* K1	*Methanohalophilus mahii* DSM 5219	*Acidaminococcus intestini* RyC-MR95	*Nostoc punctiforme* PCC 73102
Babesia bovis T2Bo	*Archaeoglobus profundus* DSM 5631	*Methanoplanus petrolearius* DSM 11571	*Acidobacterium capsulatum* ATCC 51196	*Peptostreptococcus anaerobius* 653-L
Caenorhabditis elegans	*Caldivirga maquilingensis* IC-167	*Methanopyrus kandleri* AV19	*Anaerolinea thermophila* UNI-1	*Planctomyces limnophilus* DSM 3776
Chlamydomonas reinhardtii	*Cenarchaeum symbiosum* A	*Methanoregula boonei* 6A8	*Bifidobacterium breve* DSM 20213	*Prochlorococcus marinus* str. MIT 9515
Cryptococcus neoformans var. *grubii*	*Desulfurococcus mucosus* DSM 2162	*Methanosaeta concilii* GP6	*Bulleidia extructa* W1219	*Rhodothermus marinus* SG0.5JP17-172
Cryptosporidium parvum Iowa II	*Ferroglobus placidus* DSM 10642	*Methanosalsum zhilinae* DSM 4017	*Chlamydia trachomatis*	*Spiroplasma melliferum* KC3
Dictyostelium discoideum	*Ferroplasma acidarmanus* fer1	*Methanosarcina mazei* Go1	*Chloracidobacterium thermophilum* B	*Staphylococcus aureus* A6300
Encephalitozoon cuniculi GB-M1	*Haladaptatus paucihalophilus* DX253	*Methanosphaera stadtmanae* DSM 3091	*Chloroflexus aggregans* DSM 9485	*Thermodesulfatator indicus* DSM 15286
Gallus gallus	*Halalkalicoccus jeotgali* B3	*Methanospirillum hungatei* JF-1	*Dehalogenimonas lykanthroporepellens* BL-DC-9	*Thermomicrobium roseum* DSM 5159
Giardia lamblia ATCC 50803	*Haloarcula hispanica* ATCC 33960	*Methanothermobacter thermautotrophicus* str. Delta H	*Denitrovibrio acetiphilus* DSM 12809	*Thermotoga thermarum* DSM 5069
Guillardia theta	*Halobacterium sp.* NRC-1	*Methanothermococcus okinawensis* IH1	*Desulfurispirillum indicum* S5	*Trichodesmium erythraeum* IMS101
Hydra magnipapillata	*Haloferax volcanii* DS2	*Methanothermus fervidus* DSM 2088	*Dictyoglomus turgidum* DSM 6724	*Verrucomicrobium spinosum* DSM 4136
Kluyveromyces lactis	*Halogeometricum borinquense* DSM 11551	*Methanotorris igneus* Kol 5	*Elusimicrobium minutum* Pei191	*Victivallis vadensis* ATCC BAA-548
Leishmania braziliensis MHOM	*Halomicrobium mukohataei* DSM 12286	*Natrialba magadii* ATCC 43099	*Fibrobacter succinogenes* subsp. *succinogenes* S85	

Leishmania infantum JPCM5	*Halopiger xanaduensis* SH-6	*Natrinema pellirubrum* DSM 15624	*Gemmatimonas aurantiaca* T-27
Leishmania major	*Haloquadratum walsbyi* DSM 16790	*Natronobacterium gregoryi* SP2	*Geobacter metallireducens* GS-15
Neurospora crassa	*Halorhabdus tiamatea* SARL4B	*Natronomonas pharaonis* DSM 2160	*Gloeobacter violaceus* PCC 7421
Ostreococcus lucimarinus CCE9901	*Halorubrum lacusprofundi* ATCC 49239	*Nitrosoarchaeum koreensis* MY1	*Hydrogenobacter thermophilus* TK-6
Paramecium tetraurelia strain d4-2	*Haloterrigena turkmenica* DSM 5511	*Nitrosopumilus maritimus* SCM1	*Idiomarina baltica* OS145
Physcomitrella patens	*Hyperthermus butylicus* DSM 5456	*Picrophilus torridus* DSM 9790	*Ktedonobacter racemifer* DSM 44963
Plasmodium falciparum 3D7	*Ignicoccus hospitalis* KIN4/I	*Pyrobaculum calidifontis* JCM 11548	*Leptospira interrogans serovar* Lai str. 56601
Plasmodium yoelii yoelii str. 17XNL	*Ignisphaera aggregans* DSM 17230	*Pyrococcus horikoshii* OT3	*Mariprofundus ferrooxydans* PV-1
Schizosaccharomyces pombe	*Korarchaeum cryptofilum* OPF8	*Pyrolobus fumarii* 1A	*Mesorhizobium loti* MAFF303099
Tetrahymena thermophila	*Metallosphaera cuprina* Ar-4	*Staphylothermus hellenicus* DSM 12710	*Neisseria elongata* subsp. *glycolytica* ATCC 29315
Theileria parva	*Methanobacterium sp.* AL-21	*Sulfolobus islandicus* L.S.2.15	
Trichomonas vaginalis G3	*Methanobrevibacter smithii* DSM 2375	*Thermococcus gammatolerans* EJ3	
Trypanosoma brucei	*methanocaldococcus infernus* ME	*Thermofilum pendens* Hrk 5	
Ustilago maydis	*Methanocella paludicola* SANAE	*Thermoplasma acidophilum* DSM 1728	
Zea mays	*Methanococcoides burtonii* DSM 6242	*Thermoproteus neutrophilus* V24Sta	
	Methanococcus aeolicus Nankai-3	*Thermosphaera aggregans* DSM 11486	
	Methanocorpusculum labreanum Z	*Vulcanisaeta distributa* DSM 14429	
	Methanoculleus marisnigri JR1		

Supplemental Figures

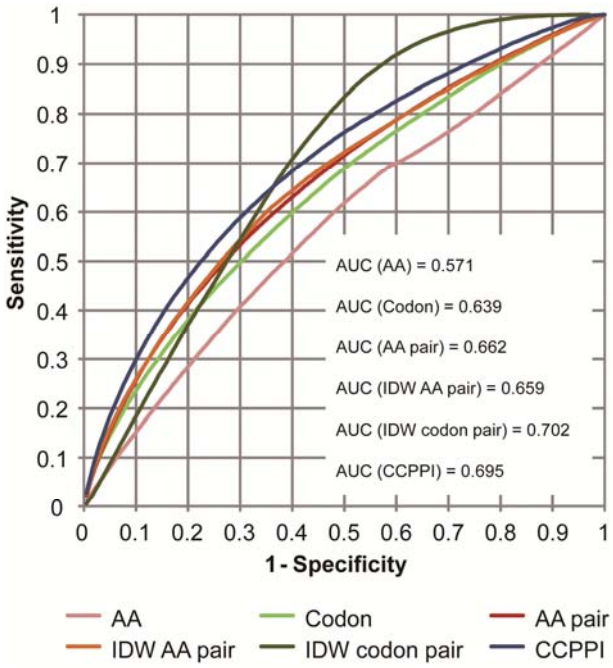

Fig. S1 The ROC curves illustrating the overall performance of the CCPPI and the other frequency difference-based encoding schemes on the large-scale testing dataset composed of the BIOGRID positives and 0.9 million random negatives. All of the encodings were trained using the "DIP+MIPS+Random" dataset and the preliminarily optimized parameters. AA here stands for amino acid.

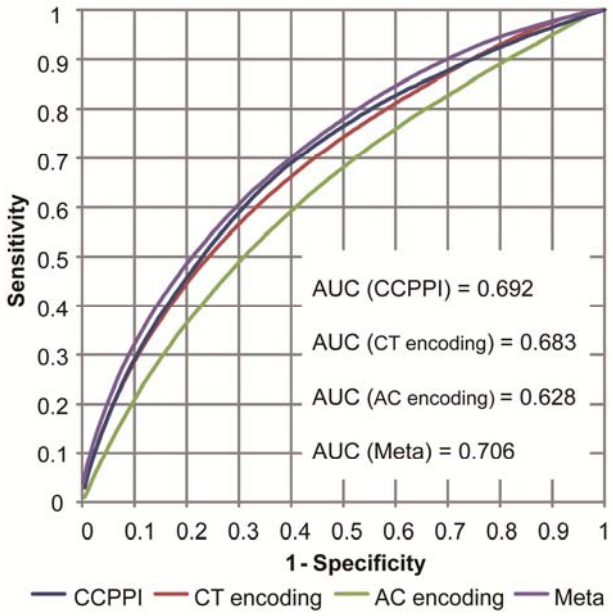

Fig. S2 The ROC curves illustrating the overall performance of the CCPPI and the other two encoding schemes on the large-scale testing dataset composed of the BIOGRID positives and 0.9 million random negatives. All three encoding schemes were trained with the "DIP+MIPS+Random" dataset and the optimized parameters listed in Table S2. The meta-predictor was constructed by weighted summing of the decision value from each predictor. The weightings of CCPPI, CT encoding and AC encoding were 1, 0.9 and 0.2, respectively.

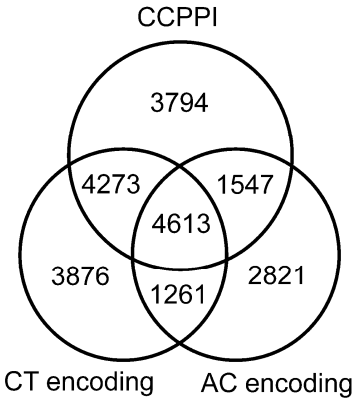

Fig. S3 Venn diagram showing the overlap of the predicted true positives by CCPPI and the other two encoding schemes at the 90% specificity level in the large-scale testing. The optimized parameters listed in Table S2 were used for SVM training of the three encoding schemes.

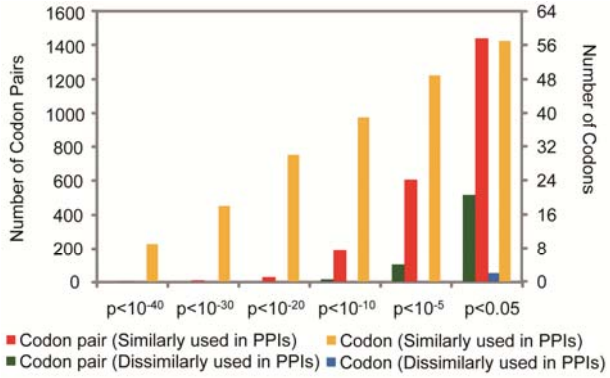

Fig. S4 Comparison of codon/codon pair frequency differences between PPIs and random ones. Redundant sequences in the datasets were removed by utilizing the CD-HIT tool to cluster sequences at 40% sequence identity cutoff. Compared with Fig. 1, the corrected p-values became less prominent (paired Wilcox's test, $p<1\times10^{-6}$). But the conclusion that many codons and codon pairs are non-randomly used in PPIs remained unchanged.

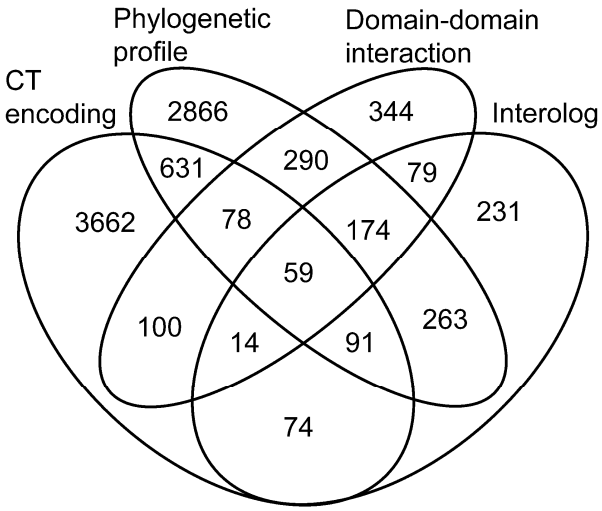

Fig. S5 Venn diagram showing the overlap of the predicted true positives from the fruit fly dataset by CT encoding and three homology-dependent PPI prediction methods. See Supplemental Methods for details of these homology-dependent methods.

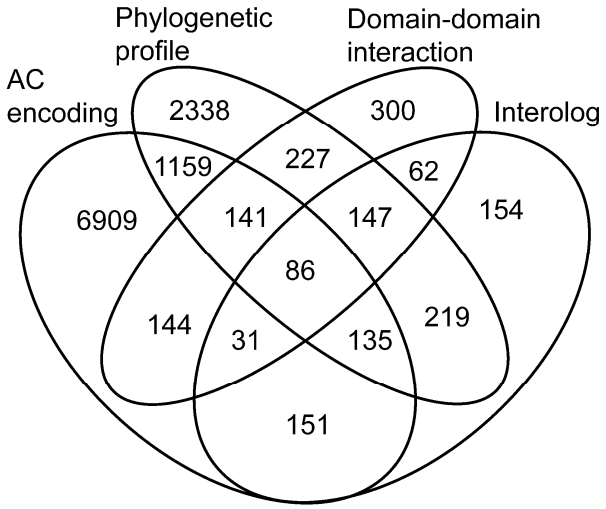

Fig. S6 Venn diagram showing the overlap of the predicted true positives from the fruit fly dataset by AC encoding and three homology-dependent PPI prediction methods. See Supplemental Methods for details of these homology-dependent methods.

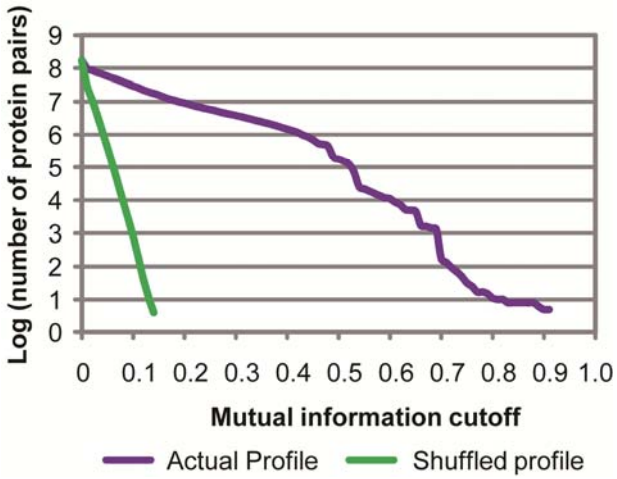

Fig. S7 The cumulative distribution of mutual information among pairs of actual and shuffled phylogenetic profiles of the fruit fly proteins.

References

1 I. Xenarios, L. Salwinski, X. J. Duan, P. Higney, S. M. Kim and D. Eisenberg, *Nucleic Acids Res.*, 2002, **30**, 303-305.
2 T. Pramila, W. Wu, S. Miles, W. S. Noble and L. L. Breeden, *Genes Dev.*, 2006, **20**, 2266-2278.
3 M. R. Flory, H. Lee, R. Bonneau, P. Mallick, K. Serikawa, D. R. Morris and R. Aebersold, *Proteomics*, 2006, **6**, 6146-6157.
4 X. Wu, L. Zhu, J. Guo, D. Y. Zhang and K. Lin, *Nucleic Acids Res.*, 2006, **34**, 2137-2150.
5 J. Shen, J. Zhang, X. Luo, W. Zhu, K. Yu, K. Chen, Y. Li and H. Jiang, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**, 4337-4341.
6 Y. Guo, L. Yu, Z. Wen and M. Li, *Nucleic Acids Res.*, 2008, **36**, 3025-3030.
7 P. J. Kersey, D. Lawson, E. Birney, P. S. Derwent, M. Haimel, J. Herrero, S. Keenan, A. Kerhornou, G. Koscielny, A. Kahari, R. J. Kinsella, E. Kulesha, U. Maheswari, K. Megy, M. Nuhn, G. Proctor, D. Staines, F. Valentin, A. J. Vilella and A. Yates, *Nucleic Acids Res.*, 2010, **38**, D563-D569.
8 A. M. Wiles, M. Doderer, J. Ruan, T. T. Gu, D. Ravi, B. Blackman and A. J. Bishop, *BMC Syst. Biol.*, 2010, **4**, 36.
9 R. D. Finn, M. Marshall and A. Bateman, *Bioinformatics*, 2005, **21**, 410-412.
10 J. Sun, J. Xu, Z. Liu, Q. Liu, A. Zhao, T. Shi and Y. Li, *Bioinformatics*, 2005, **21**, 3409-3415.
11 H. S. Najafabadi and R. Salavati, *Genome Biol.*, 2008, **9**, R87.