THE COMPARATIVE GENOMICS OF PROTEIN INTERACTIONS

JOSÉ M. PEREGRÍN-ALVAREZ[1,2] CHRISTOS A. OUZOUNIS[3]
peregrin@ebi.ac.uk ouzounis@ebi.ac.uk

[1]*Sick Kids Research Institute, TMDT MARS Building, 101 College St., 15th Floor, East Tower, M5G 1L7 Toronto, ON, Canada.*
[2]*Dpt. Of Molecular Biology and Biochemistry, University of Malaga, 29071 Malaga, Spain.*
[3]*Institute of Agrobiotechnology, Hellas 6th Km Charilaou, Thermi Rd, P.O. Box 361, 570 01 Thermi, Thessaloniki, Greece.*

The detection of gene fusion events across genomes can be used for the prediction of functional associations of proteins, including physical interactions or complex formation. These predictions are obtained by the detection of similarity for pairs of 'component' proteins to 'composite' proteins. Since the amount of composite proteins is limited in nature, we augment this set by creating artificial fusion proteins from experimentally determined protein interacting pairs. The goal is to study the extent of protein interaction partners with increasing phylogenetic distance, using an automated method. We have thus detected component pairs within seven entire genome sequences of similar size, using artificially generated composite proteins that have been shown to interact experimentally. Our results indicate that protein interactions are not conserved over large phylogenetic distances. In addition, we provide a set of predictions for functionally associated proteins across seven species using experimental information and demonstrate the applicability of fusion analysis for the comparative genomics of protein interactions.

Keywords: comparative genomics, protein interactions, *Escherichia coli*, *Helicobacter pylori*, two-hybrid screening.

1. Introduction

It has been shown that it is possible to predict protein interactions or, more generally, functional associations of proteins, including physical interaction or complex formation, using genome sequence analysis [1-3]. Fused genes encoding a single multifunctional protein in one species tend to be found in other species as pairs of genes encoding proteins showing similar functions or forming protein complexes [4]. Gene fusion is a well-known process in molecular evolution [5]; consequently, computational methods were developed to determine gene fusion events in complete genomes aiming to predict functional associations of proteins [1]. Many of these gene fusion events appear to be selectively advantageous by decreasing the regulational load in the cell for a particular process [1,3,5]. Thus, the detection of fused genes in one genome (defined as 'composite' proteins) allows the prediction of functional associations between homologous genes that remain separate in another genome (defined as 'component' proteins) [6].

Although gene fusion events (composite proteins) appear to be relatively rare [6], the accurate detection of a gene fusion event in one genome allows interactions to be predicted between many proteins across other genomes [1]. It is this kind of one-to-many relationship what makes this concept unique for discovering possible interactions or functional associations between proteins, even for those of unknown function, using

comparative genomics. Unlike other methods that rely on gene proximity to predict functional coupling [7], the gene fusion method can also detect functional relationships of distal genes within a genome. Furthermore, we have previously demonstrated the high precision of the gene fusion method using the DIFFUSE algorithm [1] (see methods and Figure 1 for a flowchart of the algorithm), which with an additional constraint of minimum alignment overlap [6] has increased to over 86% (see Methods). This computational method is analogous and complementary to the experimental approaches for the detection of protein interactions [8].

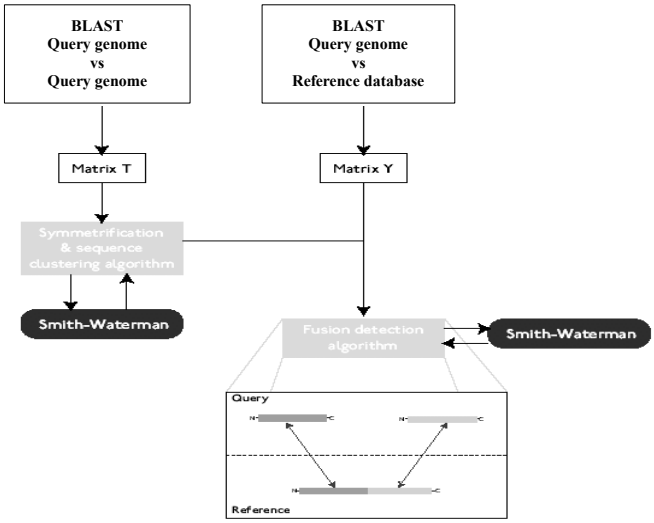

Figure 1. Flowchart of the Diffuse algorithm [1]. Similarities within the query genome using BLAST [12] are stored in a matrix T. An additional Smith-Waterman comparison [22] is used to resolve false negatives. The query genome is then compared to the reference genome, and similarities are stored in matrix Y. The fusion-detection algorithm identifies cases of the form depicted in the inset, where query proteins A and B exhibit similarity to reference protein C by checking matrix Y, but not to each other by checking matrix T, which is further confirmed by an additional Smith-Waterman comparison. Both Smith-Waterman runs are executed an additional 25 times, with randomization of the sequences, and a Z-score metric is obtained: if the Z-score is higher than certain threshold, the similarity is accepted as significant.

To examine the phylogenetic distribution (i.e conservation) of experimentally obtained protein interactions and generate predictions using comparative genomics, we have performed fusion analysis for seven entire genome sequences of similar size. Using DIFFUSE, we asked the question whether pairs of interacting proteins in *Helicobacter pylori* [9] are conserved across seven species of increasing phylogenetic distance (see methods). To achieve this, all pairs of interacting proteins from *H. pylori* have been merged to create a set of artificial fusion proteins. We define the genome where we seek component proteins as the 'query' genome and the set of artificially merged sequences from which we obtain artificial composite proteins as the 'reference' genome [1,6]. An 'artificial fusion event' is therefore defined as any pair of component proteins from a

query genome that are detected as a fused, artificial composite protein in a reference genome (Figure 1).

DIFFUSE was applied individually for each of the seven genomes, against the artificially merged sequences which are used as the reference set (see Methods). Paralogy in the query genome makes it difficult to determine precisely the actual number of possible associations and increases uncertainty in accurately predicting component proteins [1]. However, the detection of component pairs, in different species with similarly sized genomes, via their similarity to interacting pairs in *H. pylori* followed by and additional constraint of minimal alignment overlap, allows us to assess the extent of the conservation of interacting proteins and generate predictions of protein interactions for distantly related species.

2. Methods

To generate predictions of functional associations in complete genomes, we have extracted the information and sequences of interacting proteins (6,797 sequences involved in 11,251 protein interactions) from the Database of Interacting Proteins (DIP) [10]. Of these interactions, 13% refer to *Helicobacter pylori* and were subsequently used in this analysis. We artificially merged sequence pairs using their original DIP binary relationships in order to get a Reference data set (Figure 1).

All selected genome sequences were obtained from their original sources [11] and used as queries. Genomes were selected based on phylogenetic distance to the reference genome *H. pylori*. Phylogenetic distance was defined by phylogenetic depth in a species-tree built by using Small Subunit rRNAs from the Ribosomal Database Project (RDP) [http://rdp.cme.msu.edu/html/].

The query database is compared against itself using BLASTp [12] (E-value threshold 10^{-06}), after masking compositional biased regions with the CAST [13] algorithm (score threshold 40), and all pairwise sequence similarities are recorded in a binary matrix. The query database is also compared against the reference database, as above, and similarities are recorded in another binary matrix. The DIFFUSE algorithm [1] (see Figure 1 for a flowchart of the algorithm) was then applied to both matrices and the detected 'artificial' gene fusion results were further filtered for significant overlap by more than 10% of their total length when aligned together with the artificial composite protein [6].

To assess the quality of the interaction and prediction data, we filtered all interacting pairs using either a functional class or subcellular localization criteria. We created 3 categories for functional classes (identical or positive, different, and unknown class), and 2 categories for subcellular localization (identical or positive, and different). The information about functional classes and subcellular localization was extracted from GeneQuiz [14] and MIPS [15], respectively.

The total automatic analysis was performed over a period of 72 hours on a 4-CPU Sun E450 with 2GB of RAM.

3. Results

As a first estimate of the performance of our approach and in order to assess the quality of our predictions, we have first tested a *H. pylori*-related species, the complete genome of the Proteobacteria *Escherichia coli* (4,290 ORFs) as query, against 1,359 *Helicobacter pylori* artificially fused sequences, defined as the reference set. We thus detected possible interacting partners using the DIFFUSE algorithm [1,6] and subsequently made predictions of functional associations of proteins across species. True positive protein interactions are expected to involve protein partners that belong to the same functional class [16]. Hence, we tested whether the artificial fusion events and component proteins identified by the DIFFUSE algorithm tend to involve component proteins with similar functional annotation. Thus, whenever a possible artificial fusion event is found, the artificial composites and the two components detected are assigned to 3 classes of functional information (see Methods). Conceptually, this approach is similar to the comparative analysis of protein interactions for *H. pylori* and *E. coli* [17]. Our analysis yields 1,487 pairs of *E. coli* proteins, with 141 (9%) classified as positives cases (i.e. in the same functional class)(see methods), 711 (48%) in different classes and 635 (43%) have at least one component with no functional class assignment (unknown) (see Methods).

To enhance the quality of our predictive analysis and eliminate noise from experimental procedures, we only consider components classified in the same functional class, thus all observations of different classes or those not fully classified are not further considered. This assumption obviously decreases the predictive potential of our method because it ignores cases of pairs of interacting proteins without class assignment, which may be functionally related [1,16]. The detected component proteins are far fewer in number than for the five previous reports of *E. coli* [1,3,6,16,17], due to the much more stricter criteria employed in this study and the multi-step protocol we have developed. Of these 141 positive cases only 12 appear to represent the same pairs of interacting proteins (putative orthologs)(see methods) in both the query (components) and reference set (artificial composites) (Table 1). Our method identifies a number of well-known interacting protein pairs. These are proteins participating in the same protein complex or biochemical process, such as Regulatory functions, Replication, Transcription, Translation and Transport-and-Binding proteins, according to the GeneQuiz functional classification (formed by 15 different functional classes) [14]. A number of unconfirmed cases constitute some interesting testable predictions. For example the phosphate regulon transcriptional regulatory protein PhoB was predicted to interact with the chemotaxis protein CheA (see URL in Discussion). The other 129 cases do not have consistent annotations across the two species because they are either paralogous genes or share specific domains with one or both of the artificial fused genes.

Coverage cannot easily be estimated, as we do not know in advance how many proteins potentially interact within the query genome. Thus, we have tried to estimate coverage by using the *H. pylori* genome as query against the *H. pylori* artificially fused

sequences reference set, as described above, and then counting the number of predicted artificial composites.

Table 1. Prediction of protein interactions in *E. coli* using artificially fused *H. pilory* sequences.

S	FC	SI	ID	FUNCTION	SI	ID	FUNCTION
R	**Rg**	HP1067	CHEY_HELPY	Chemotaxis protein CheY	HP0392	O25153	Chemotaxis protein CHEA (EC 2.7.3.-)
P		1788191	CHEY_ECOLI		1788197	CHEA_ECOLI	
R	**Rp**	HP0705	UVRA_HELPY	Excinuclease ABC subunit A	HP1541	MFD_HELPY	Transcription-repair coupling factor (TRCF)
P		2367343	UVRA_ECOLI		1787357	MFD_ECOLI	
R	**Rp**	HP0705	UVRA_HELPY	Excinuclease ABC subunit A	HP1114	UVRB_HELPY	Excinuclease ABC subunit B
P		2367343	UVRA_ECOLI		1786996	UVRB_ECOLI	
R	**Rp**	HP0705	UVRA_HELPY	Excinuclease ABC subunit A	HP0821	UVRC_HELPY	Excinuclease ABC subunit C
P		2367343	UVRA_ECOLI		1788221	UVRC_ECOLI	
R	**Tc**	HP1293	RPOA_HELPY	RNA polymerase alpha subunit	HP1198	O25806	RNA polymerase beta chain (EC 2.7.7.6)
P		1789690	RPOA_ECOLI		1790419	RPOB_ECOLI	
R	**Tl**	HP0399	RS1_HELPY	30S ribosomal protein S1	HP1048	IF2_HELPY	Translation initiation factor IF-2
P		1787140	RS1_ECOLI		1789559	IF2_ECOLI	
R	**Tl**	HP1246	RS6_HELPY	30S ribosomal protein S6	HP1244	RS18_HELPY	30S ribosomal protein S18
P		1790644	RS6_ECOLI		1790646	RS18_ECOLI	
R	**Tl**	HP1246	RS6_HELPY	30S ribosomal protein S6	HP0886	SYC_HELPY	Cysteinyl-tRNA synthetase (EC 6.1.1.16)
P		1790644	RS6_ECOLI		1786737	SYC_ECOLI	
R	**Tl**	HP1312	RL16_HELPY	50S ribosomal protein L16	HP0083	RS9_HELPY	30S ribosomal protein S9 (BS10)
P		1789709	RL16_ECOLI		1789625	RS9_ECOLI	
R	**Tl**	HP1312	RL16_HELPY	50S ribosomal protein L16	HP1316	RL2_HELPY	50S ribosomal protein L2
P		1789709	RL16_ECOLI		1789713	RL2_ECOLI	
R	**Tp**	HP0687	O25396	Ferrous iron transport protein B	HP1072	COA0_HELPY	Copper-transporting ATPase (EC 3.6.1.36)
P		1789813	FEOB_ECOLI		1786691	ATCU_ECOLI	
R	**Tp**	HP0687	O25396	Ferrous iron transport protein B	HP1506	O26036	Sodium/Glutamate symport carrier protein
P		1789813	FEOB_ECOLI		1790085	GLTS_ECOLI	

The 12 positive pairs with identical functional class assignments (see Methods): pairs of interacting proteins in the query (*E. coli* components) and reference set (*H. pylori* artificial composites). Column names: Source (S) of information divided into Reference set (R) and the prediction for *E. coli* (P); Functional Class (FC) (Rg, Regulatory functions; Rp, Replication; Tc, Transcription; Tl, Translation; and Tp, Transport-and-binding proteins); Identifier (ID) from Swissprot and Fuctional assignment (Function), Sequence Identifier (SI), according to Genequiz [14]. Table is sorted by Functional class. Empty cells in Function columns, for simplicity, imply identical assignment between Reference and Prediction. Columns 3-5 and 6-8 correspond to the details of the individual component proteins. Predictions were performed with the DIFFUSE algorithm [1,6] using the *H. pylori* artificially merged sequences as Reference set and the *E. coli* genome as query (see Figure 1).

This number is equal to the number of composites in the original reference set, i .e. we obtained a 100% coverage when self-interactions are removed from the original DIP [10] data set, and 96% coverage when self-interactions are not removed, since DIFFUSE cannot detect sequence-similar components (see methods) [1,6]. We then calculated the

percentage of interacting proteins (components) that are shared by the two species, defined as the number of unique (non-redundant) detected components in *E. coli* vs. the *H. pylori* reference set, divided by the number of unique components in a control experiment of the *H. pylori* genome as query vs. *H. pylori* reference set. This yielded a 60% of conserved protein pairs. This number strongly depends on the phylogenetic distance between query and reference genomes [6] but, a priori, it suggests that protein interactions (as pairs of proteins) may not be strongly conserved across related species.

To investigate the conservation of interacting pairs across other genomes in a more consistent way, we have repeated this analysis across other six species with similar genome sizes. Previously, we have explored the influence of three key factors in gene fusion analysis: genome size, paralogy and phylogenetic distance [6]. Herein, we focus on the latter, to understand the conservation of protein interactions from a comparative genomics perspective. To investigate further the results obtained from the comparison with *E. coli* (see above), we used an additional six genomes of similar size: *Campylobacter jejuni* (1,634 ORFs), *Haemophilus influenzae* (1,707 ORFs), *Borrelia burgdorferi* (1,639 ORFs), *Streptococcus pyogenes* (1,696 ORFs), *Thermotoga maritima* (1,849 ORFs) and *Thermoplasma acidophilum* (1,478 ORFs); plus the *H. pylori* genome (1,575 ORFs) as control. Since the number of components detected in each species depends on genome size, we have selected genomes according to two criteria: roughly similar genome size as the *H. pylori* genome (ranging between 1,478 and 1,849 ORFs) and a wide range of phylogenetic distances to *H. pylori* (Figure 2). As expected, there is an inverse relationship between the number of components and phylogenetic distance. This trend is only violated in the case of the *S. pyogenes* and *T. maritima* genomes, partly explained by the degree of paralogy for certain proteins [1]. The key conclusion from this analysis is that interactions as 'pairs' of protein partners seem are not highly conserved, are eroded over large phylogenetic distances, and may correspond to species-specific instances of interacting pairs (Figure 2).

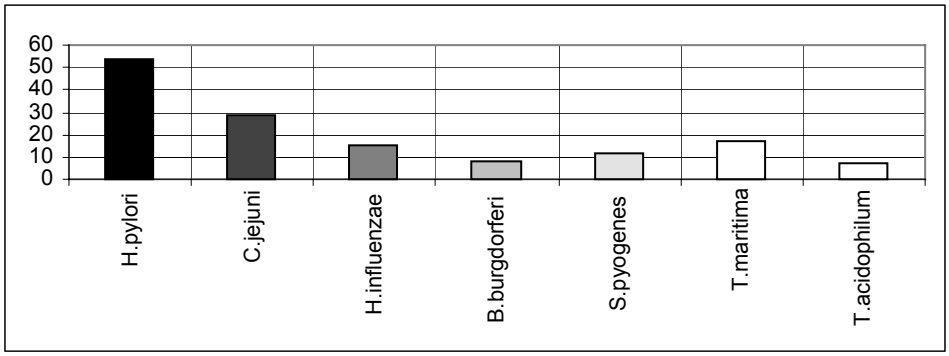

Figure 2. Numbers of protein interactions detected in seven genomes. Genomes are shown on the x-axis and the relative number of artificial component pairs (putative homolog interactions) on the y-axis. Genomes are sorted by decreased phylogenetic distance (see methods) to the *H. pylori* genome, which was used as control set. Bar shading from black to white represents phylogenetic distance to *H. pylori (*last two bars with white color for simplicity). Counts are normalized by genome size, although genomes sizes are comparable (actual counts are shown in Table 2A, row: TOTAL).

How many of the conserved pairs are predicted to belong to the positive class (i.e. same functional class)? When we perform the same analysis using functional class

assignments, it is possible to associate those with phylogenetic distance (Table 2A). Again, generally speaking, the trend of decreasing number of interactions over large phylogenetic distances is observed, although less clearly – due to various degrees of annotation accuracy, obtained from GeneQuiz [14]. We again focus on positive cases (see methods) where predictions are of the highest quality (Table 2B). Among all predicted positive cases only predictions assigned to the Transcription class, namely between RNA polymerase subunits, are detected in 6 out of the 7 genomes. The method cannot predict the same pair of interacting proteins in the *Thermoplasma acidophilum* genome, due to the highly divergent nature of the archaeal RNA polymerase. This fact further illustrates the point that only a few interactions seem to be highly conserved or detectable across large evolutionary distances (see discussion).

Table 2. Prediction of protein interactions in complete genomes. Predictions were performed by the DIFFUSE algorithm (see Figure 1) [1] using the *H. pylori* artificially fused set as Reference and 6 different genomes as queries. **(A)** Number of components pairs in different genomes. The *H. pylori* genome was used as a control. Column names: Categories according to their distribution of functional classes (see Methods) followed by species names (*). Species (columns 2 to 8) are sorted by phylogenetic distance to *H. pylori.* **(B)** Potential positive examples of pairs of interacting proteins in the query genomes (components) and reference set (artificial composites) are listed.

(A)

Categories	***H. pylori***	***C. jejuni***	***H. influenzae***	***B. burgdorferi***	***S. pyogenes***	***T. maritima***	***T. acidophilum***
Same functional class	41	48	39	14	26	30	15
Different functional class	292	300	284	54	293	173	101
At least one unknown functonal class	4461	935	196	195	285	600	70
TOTAL	**4794**	**1283**	**519**	**263**	**604**	**803**	**186**

(B)

S	**FC**	**SI**	**ID**	**FUNCTION**	**SI**	**ID**	**FUNCTION**
R	**Tc**	HP1293	RPOA_HELPY	RNA polymerase alpha	HP1198	O25806	RNA polymerase beta
Hp	**Tc**	HP1293	RPOA_HELPY		HP1198	O25806	
Cj	**Tc**	6969012	RPOA_CAMJE		6967949	RPOB_CAMJE	
Hi	**Tc**	HI0802	RPOA_HAEIN		HI0515	RPOB_HAEIN	
Bb	**Tc**	BB0502	RPOA_BORBU		BB0389	RPOB_BORBU	
Sp	**Tc**	13621394	RPOA_STRPY		13621404	Q9A1U1	
Tm	**Tc**	TM1472	RPOA_THEMA		TM0458	RPOB_THEMA	
Cj	**Tc**	6969012	RPOA_CAMJE	RNA polymerase alpha	6967950	RPOC_CAMJE	RNA polymerase beta'
Hi	**Tc**	HI0802	RPOA_HAEIN		HI0514	RPOC_HAEIN	
Bb	**Tc**	BB0502	RPOA_BORBU		BB0388	RPOC_BORBU	
Sp	**Tc**	13621394	RPOA_STRPY		13621405	RPOC_STRPY	
Tm	**Tc**	TM1472	RPOA_THEMA		TM0459	RPOC_THEMA	
Cj	**Tc**	6967949	RPOB_CAMJE	RNA polymerase beta	6967950	RPOC_CAMJE	RNA polymerase beta'
Hi	**Tc**	HI0515	RPOB_HAEIN		HI0514	RPOC_HAEIN	
Bb	**Tc**	BB0389	RPOB_BORBU		BB0388	RPOC_BORBU	
Sp	**Tc**	13621404	Q9A1U1		13621405	RPOC_STRPY	
Tm	**Tc**	TM0458	RPOB_THEMA		TM0459	RPOC_THEMA	
Ta	**Tc**	10639562	RPOB_THEAC	RNA polymerase B	10639564	RPA2_THEAC	RNA polymerase A"

Column names: as in Table 1. Note that only the first case corresponds to the reference set, while the other cases are identified due to paralogy between the corresponding components. (*) Abbreviations: Hp, *Helicobacter pylori*; Cj, *Campylobacter jejuni*; Hi,*Haemophilus influenzae*; Bb, *Borrelia burgdorferi*; Sp, *Streptococcus pyogenes*; Tm, *Thermotoga maritima*; and Ta, *Thermoplasma acidophilum*.. Tc, Transcription class.

In order to assess the quality of the original data from DIP [10] in terms of predicting functional associations, we analyzed the patterns of distribution of Functional Classes for protein pairs (in terms of positive, different or unknown classes)(see methods) according to three different annotation schemes: Clusters of Orthologous Groups (COGs) [18], Euclid [19] and GeneQuiz [14]. The COG scheme yields more cases in the same functional class compared to the other two schemes in relative terms (9% of interactions in the reference set), although only 1106 sequences out of the total 1575 sequences from *H. pylori* genome are assigned to COGs. Therefore, we opted using the GeneQuiz functional class scheme (2%), since it provides maximum coverage of the genome, similarly to Euclid (3%) [19].

We then assessed the patterns of distribution per functional class in the positive category (see methods) for the six selected genomes and the reference set (Figure 3). In general, functional class distribution of positive pairs exhibits a highly non-uniform pattern for all six species examined and the reference set. The three more abundant functional classes correspond to the transport-and-binding-proteins, replication and translation, in this order, whose pattern is different from the reference set, suggesting that the observed interactions in *H. pylori* are not conserved in terms of functional class. The cell envelope functional class, which is one of the most abundant classes in the reference set, does not give rise to any predictions across the selected six genomes except for the case of the *H. pylori* genome used as control.

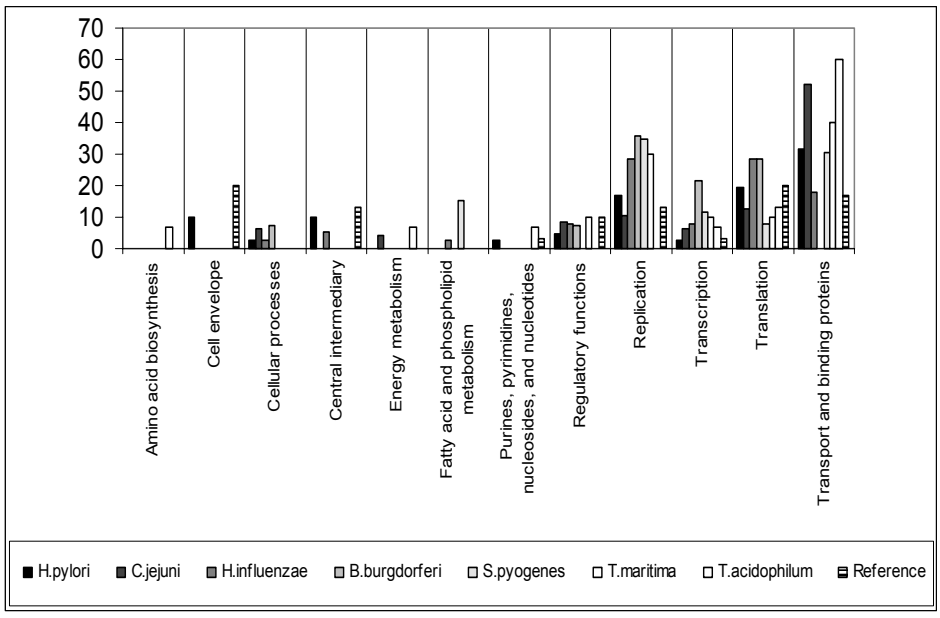

Figure 3. Relative distribution of functional classes for pairs of component proteins in the positive category. Functional classes are shown on the x-axis and the number of component pairs (as %) per functional class on the y-axis. Genomes are sorted by phylogenetic distance to the *H. pylori* genome, which was used as a control set. Functional class are sorted out by alphabetical order. Bar shading as in Figure 2 but the reference set which is shown as dark horizontal.

4. Discussion

Is there any bias in protein interaction information stemming from specific experimental techniques? Given the fact that all the *H. pylori* data comes exclusively from two-hybrid screening tests [9], we examine the influence of other experimental techniques for protein interaction detection. We counted the number of protein interactions per method (available in DIP), finding out that there are only two (out of 38) methods with more than 1,000 interactions per method (out of 11,251 recorded protein interactions in total). These two methods are two-hybrid screening and immunoprecipitation with 9,781 and 1,243 protein interactions, respectively. The remaining cases all have counts less than 400 interactions per method for a variety of species.

In fact, the only species (out of 112 species represented in DIP) for which there is sufficient information of protein interactions obtained by more than one method, including two-hybrid screening and immunoprecipitation is *Saccharomyces cerevisiae*. Protein interactions from this species can be validated not only by functional classes but also by subcellular localization. When all interacting proteins from *S. cerevisiae* are extracted and classified according to the corresponding experimental method, immunoprecipitation appears to yield more consistent functional classes (28% of interactions) and subcellular localization (75% of interactions), compared to two-hybrid screening (5% and 44%, respectively). Thus, there is a potential bias due to the experimental data used in our study. Furthermore, two-hybrid screening have been shown to have a high false positive (and false negative) rate [20]. Our results also show that subcellular localization is a highly desirable attribute for protein interactions that could be used both as a quality measure for the original DIP information as well as a filter for predicting reliable interactions. We should also note that only in a handful of cases, the two above mentioned experimental methods result in identical pairs of proteins, indicating that there is very little overlap of reliable experimental observations [21] (not shown).

Another caveat of our approach is that our results on the poor conservation of interactions across genomes may be due to difficulty of retrieving such relationships across large phylogenetic distances due to the use of BLAST [12] for protein similarity searches. Therefore, there is a possibility that the interaction pairs are indeed maintained across evolution but our approach failed to retrieve it. Furthermore, our approach rely on the quality and extent of the functional annotation data. Since our analysis shows that annotation schemes differ across genomes and it is biased towards identifying proteins of known function in more distantly related taxa (see Table 2) this might represent a bias in the extent and quality of the annotations used in this study.

In summary, although our approach has some caveats that will be addressed in a larger scale follow-up analysis, our preliminary analysis shows that the exhaustive detection of 'artificial' gene fusion events allows the prediction of functionally associated components based merely on genome structure. This approach for the prediction of functional associations of proteins results in accurate predictions for physical interactions, pathway

involvement, complex formation and other types of functional associations of protein molecules, many of them may provide further support for previous studies [1,3,6,16,17].

All the cases presented here are based on the high-throughput protein interaction set from *H. pylori* and represent interesting and novel findings available at the following URL: http://cgg.ebi.ac.uk/old/cgg//projects/mining/artifuse/.

Acknowledgments

We thank members of the former Computational Genomics Group for discussions. This work was supported by the European Molecular Biology Laboratory and the Ministry of Science and Technology, Spain. C. A. O. thanks the UK Medical Research Council and IBM Research for additional support.

References

[1] Enright, A.J., Iliopoulos, I., Kyrpides, N.C. and Ouzounis, C.A. (1999) Nature 402, 86-90.
[2] Marcotte, E.M., Pellegrini, M., Thompson, M.J., Yeates, T.O. and Eisenberg, D. (1999) Nature 402, 83-6.
[3] Marcotte, E.M., Pellegrini, M., Ng, H.L., Rice, D.W., Yeates, T.O. and Eisenberg, D. (1999) Science 285, 751-3.
[4] Sali, A. (1999) Nature 402, 23, 25-6.
[5] Doolittle, R.F. (1999) Nat Genet 23, 6-8.
[6] Enright, A.J. and Ouzounis, C.A. (2001) Genome Biol 2(9):research0034.1-0034.7.
[7] Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D. and Maltsev, N. (1999) Proc Natl Acad Sci U S A 96, 2896-901.
[8] Ito, T. et al. (2000) Proc Natl Acad Sci U S A 97, 1143-7.
[9] Rain, J.C. et al. (2001) Nature 409, 211-5.
[10] Xenarios, I., Salwinski, L., Duan, X.J., Higney, P., Kim, S.M. and Eisenberg, D. (2002) Nucleic Acids Res 30, 303-5.
[11] Kyrpides, N.C. (1999) Bioinformatics 15, 773-4.
[12] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Nucleic Acids Res 25, 3389-402.
[13] Promponas, V.J., Enright, A.J., Tsoka, S., Kreil, D.P., Leroy, C., Hamodrakas, S., Sander, C. and Ouzounis, C.A. (2000) Bioinformatics 16, 915-22.
[14] Andrade, M.A. et al. (1999) Bioinformatics 15, 391-412.
[15] Mewes, H.W. et al. (2002) Nucleic Acids Res 30, 31-4.
[16] Yanai, I., Derti, A. and DeLisi, C. (2001) Proc Natl Acad Sci U S A 98, 7940-5.
[17] Wojcik, J. and Schachter, V. (2001) Bioinformatics 17, S296-305.
[18] Tatusov, R.L. et al. (2001) Nucleic Acids Res 29, 22-8.
[19] Tamames, J., Ouzounis, C., Casari, G., Sander, C. and Valencia, A. (1998) Bioinformatics 14, 542-3.
[20] Deane C., Salwiński Ł., Xenarios I., Eisenberg D. (2002). Mol Cell Proteomics **1** (5): 349-56.

[21] von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S. and Bork, P. (2002) Nature 417, 399-403.
[22] Smith T.F. & Waterman M. J. (1981). J Mol Biol 147, 195-197.