BMC Bioinformatics

Methodology article

Domain fusion analysis by applying relational algebra to protein sequence and domain databases

Kevin Truong and Mitsuhiko Ikura*

Address: Department of Medical Biophysics, University of Toronto, Toronto, M3N 1L6, Canada

Email: Kevin Truong - ktruong@uhnres.utoronto.ca; Mitsuhiko Ikura* - mikura@uhnres.utoronto.ca

* Corresponding author

Published: 6 May 2003

BMC Bioinformatics 2003, **4**:16

Received: 25 February 2003
Accepted: 6 May 2003

This article is available from: http://www.biomedcentral.com/1471-2105/4/16

Abstract

Background: Domain fusion analysis is a useful method to predict functionally linked proteins that may be involved in direct protein-protein interactions or in the same metabolic or signaling pathway. As separate domain databases like BLOCKS, PROSITE, Pfam, SMART, PRINTS-S, ProDom, TIGRFAMs, and amalgamated domain databases like InterPro continue to grow in size and quality, a computational method to perform domain fusion analysis that leverages on these efforts will become increasingly powerful.

Results: This paper proposes a computational method employing relational algebra to find domain fusions in protein sequence databases. The feasibility of this method was illustrated on the SWISS-PROT+TrEMBL sequence database using domain predictions from the Pfam HMM (hidden Markov model) database. We identified 235 and 189 putative functionally linked protein partners in *H. sapiens* and *S. cerevisiae*, respectively. From scientific literature, we were able to confirm many of these functional linkages, while the remainder offer testable experimental hypothesis. Results can be viewed at http://calcium.uhnres.utoronto.ca/pi.

Conclusion: As the analysis can be computed quickly on any relational database that supports standard SQL (structured query language), it can be dynamically updated along with the sequence and domain databases, thereby improving the quality of predictions over time.

Background

The complex metabolic and signaling pathways within the cell are controlled by highly coordinated and intricate protein-protein interactions. Information regarding such protein-protein interactions can be obtained from biochemical and biophysical methods like co-immunoprecipitation [1], yeast two-hybrid [2] and mass spectrometry [3,4]. To complement these often time-consuming experimental methods, computational methods for predicting functional linkages have been developed. Some methods use protein surface interfaces [5,6]; some use the ordering and/or proximity of genes in genomes [7–9]; while others use the co-occurrences of genes in genomes [10,11].

Recently, computational methods that exploit domain-domain relationships have been introduced and proven to be useful for the prediction of functional linkages in genomic research [12–15]. In particular, domain fusion analysis exploits the fact that certain proteins in a given genome consist of fused domains that correspond to single, full-length proteins in other genomes [13–15]. The proteins with fused domains in a given genome are likely to directly interact or be involved in the same metabolic and signaling pathways. In their analysis of the *M. genitalium* genome, Huynen *et al.* showed that the occurrence of a domain fusion event was highly correlated with function [16].

The query genome is defined as the genome where functional linkages are predicted, while the reference genome is the amalgamation of all other genomes excluding the query. Domain fusion events found in the reference genome predict functional linkages between proteins in the query. To date, most domain fusion analysis have compared complete genomes of relatively small sizes and rely on a BLAST comparison [17] between every protein of the query genome to every protein of the reference [13,14,16,18–20]. The analysis has not been applied to larger non-redundant sequence databases such as SWISS-PROT, although the analysis becomes a more powerful prediction tool when more reference genomes are included. One reason for this limitation is that the computation time becomes "prohibitively expensive" [18].

Other groups have already appreciated the use of relational databases for domain fusion analysis [21,22]. To complement their work, we present a fast computational method that enables domain fusion analysis on partial or complete genomes in a non-redundant sequence database using simple relational algebra operations. Instead of using BLAST comparisons, we leveraged on existing efforts to predict protein domains by Pfam's HMM domain database [23]. Beginning with Pfam's domain layout prediction of each protein in the SWISS-PROT+TrEMBL protein sequence database, we applied successive relational algebra operations using SQL to identify putative functional linkages, especially in *H. sapiens* and *S. cerevisiae*. These results are compared with experimentally demonstrated cases and published protein interaction databases. Finally, we discuss various factors that can generate false positives.

Algorithm

The majority of protein sequence and domain databases are built on the relational database architecture. Typically, data is acquired from a database of this type by relational algebra operations in the form of SQL queries. Therefore, a method that can be performed directly using these operations will save unnecessary conversion of data and leverage on the scalability and efficiency of commercial RDBMS software (relational database management systems). Our method for finding domain fusions can be performed entirely using relational algebra operations.

The method is described using relational algebra notation with the following conventions: **bold** text refers to a table; **A**.attribute refers to an attribute or column of **A**; $\sigma_{(\text{predicate})}(\mathbf{A})$ is the selection operation with the predicate in parenthesis, which selects rows in **A** that satisfy the predicate; **A** × **B** is the cartesian product operation, which creates a permutation of information between **A** and **B**; $\pi_{(\text{A.attribute1, A.attribute2,...})}(\mathbf{A})$ is the projection operation, which extracts specified attributes from **A**.

As a minimum, the database must have a sequence table (denoted by **S**) and a domain layout table (denoted by **D**) with some key attributes (Figure 1a). **D** stores the domain layout of each of the sequences in **S** and is linked by the seq_id attribute. Let $\mathbf{S}_{\mathbf{query}}$ be the table of all protein sequences in the sequence database of the query genome and let $\mathbf{S}_{\mathbf{ref}}$ be the table of all other sequences in the database comprising the reference genomes. Let $\mathbf{D}_{\mathbf{query}}$ and $\mathbf{D}_{\mathbf{ref}}$ be the tables of all domain layouts belonging to the query and reference genomes, respectively. Therefore, $\mathbf{S}_{\mathbf{query}}$, $\mathbf{S}_{\mathbf{ref}}$, $\mathbf{D}_{\mathbf{query}}$, and $\mathbf{D}_{\mathbf{ref}}$ are defined by the following relations:

$$\mathbf{S}_{\mathbf{query}} = \sigma_{\text{genome="query genome"}}(\mathbf{S}) \quad (1)$$

$$\mathbf{S}_{\mathbf{ref}} = \sigma_{\text{genome}\neq\text{"query genome"}}(\mathbf{S}) \quad (2)$$

$$\mathbf{D}_{\mathbf{query}} = \pi_{\text{D.seq\_id,D.dom}}(\sigma_{\text{D.seq\_id}=\text{S}_{\text{query}}\text{.seq\_id}}(\mathbf{D} \quad \mathbf{S}_{\mathbf{query}})) \quad (3)$$

$$\mathbf{D}_{\mathbf{ref}} = \pi_{\text{D.seq\_id,D.dom}}(\sigma_{\text{D.seq\_id}=\text{S}_{\text{ref}}\text{.seq\_id}}(\mathbf{D} \times \mathbf{S}_{\mathbf{ref}})) \quad (4)$$

Let $\mathbf{F}_{\mathbf{query}}$ and $\mathbf{F}_{\mathbf{ref}}$ be the table of all possible domain fusion templates (DFTs) in the query and reference genomes, respectively. The idea of DFTs is conceptually similar to Rosetta stone [14] and composite proteins [13]. For example, if a gene has four different domains ABCD, there are six different DFTs: AB, AC, AD, BC, BD, and CD.

Let $\mathbf{D}_{\mathbf{q1}}$ and $\mathbf{D}_{\mathbf{q2}}$ be copies of $\mathbf{D}_{\mathbf{query}}$ and let $\mathbf{D}_{\mathbf{r1}}$ and $\mathbf{D}_{\mathbf{r2}}$ be copies of $\mathbf{D}_{\mathbf{ref}}$. $\mathbf{F}_{\mathbf{query}}$ and $\mathbf{F}_{\mathbf{ref}}$ can be found by performing a projection and selection operation following a cartesian product between the corresponding domain tables. This operation will enumerate all permutations of DFTs. For example, if gene has three different domains ABC, then there are nine possible permutations of DFTs: AA, AB, AC, BA, BB, BC, CA, CB, and CC. The desired DFTs do not have the same domains (i.e., AA, BB, and CC) and order does not matter (i.e., AB is the same as BA). To remove same domain DFTs, the following clauses are added to the selection predicates: ($\mathbf{D}_{\mathbf{q1}}$.dom≠$\mathbf{D}_{\mathbf{q2}}$.dom) for $\mathbf{F}_{\mathbf{query}}$ and ($\mathbf{D}_{\mathbf{r1}}$.dom≠$\mathbf{D}_{\mathbf{r2}}$.dom) for $\mathbf{F}_{\mathbf{ref}}$. At this stage, it is not necessary to consider the removal of one of the two alternatively ordered DFTs.

$$\mathbf{F}_{\mathbf{query}} = \pi_{\text{D}_{\text{q1}}\text{.dom,D}_{\text{q2}}\text{.dom}}(\sigma_{(\text{D}_{\text{q1}}\text{.seq\_id=D}_{\text{q2}}\text{.seq\_id})\wedge(\text{D}_{\text{q1}}\text{.dom}\neq\text{D}_{\text{q2}}\text{.dom})}(\mathbf{D}_{\mathbf{q1}} \quad \mathbf{D}_{\mathbf{q2}})) \quad (5)$$

$$\mathbf{F}_{\mathbf{ref}} = \pi_{\text{D}_{\text{r1}}\text{.dom,D}_{\text{r2}}\text{.dom}}(\sigma_{(\text{D}_{\text{r1}}\text{.seq\_id=D}_{\text{r2}}\text{.seq\_id})\wedge(\text{D}_{\text{r1}}\text{.dom}\neq\text{D}_{\text{r2}}\text{.dom})}(\mathbf{D}_{\mathbf{r1}} \quad \mathbf{D}_{\mathbf{r2}})) \quad (6)$$

Let $\mathbf{F}_{\mathbf{put}}$ be the table of valid DFTs that can be used in the prediction of functionally linked proteins in the query

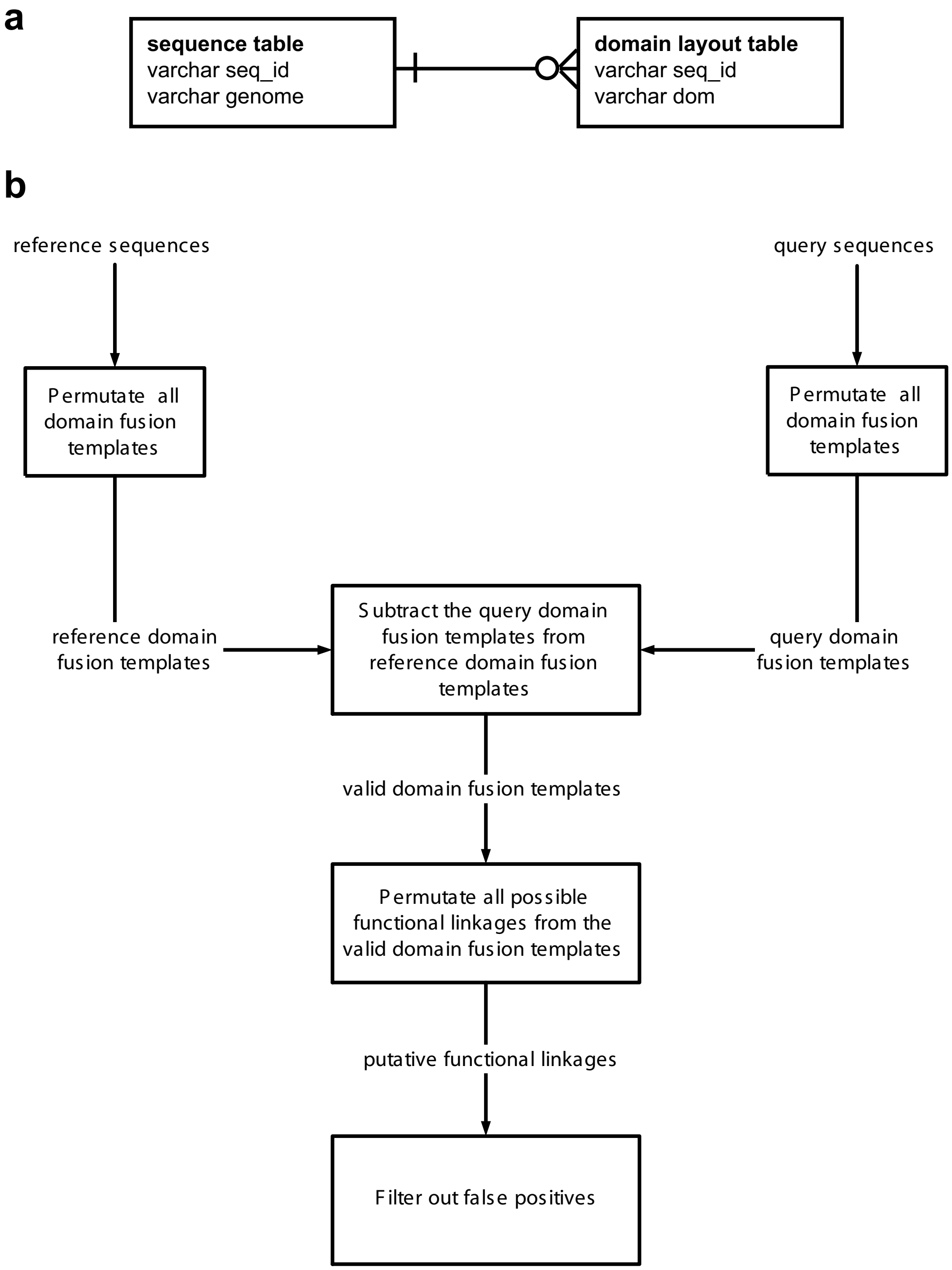
a
sequence table
varchar seq_id
varchar genome
domain layout table
varchar seq_id
varchar dom
b
reference sequences
query sequences
Permutate all domain fusion templates
Permutate all domain fusion templates
reference domain fusion templates
Subtract the query domain fusion templates from reference domain fusion templates
query domain fusion templates
valid domain fusion templates
Permutate all possible functional linkages from the valid domain fusion templates
putative functional linkages
Filter out false positives

Figure 1
(a) The Crow's Foot entity relationship diagram ofthe database architecture. (b) The flowchart of the method Protein sequences from SWISS-PROT+TrEMBL were divided into query sequences (belonging to the genome of interest) or reference sequences (everything else). All possible DFTs of the query or reference sequences were then permutated using a single SQL command. The valid DFTs were found by subtracting the query set from the reference set. Finally, using the valid DFTs, all functional linkages were permutated. The number of putative functional linkages is generally large, so it is necessary to filter out false positives.

genome. Therefore, $\mathbf{F_{put}}$ can be found by the difference between $\mathbf{F_{ref}}$ and $\mathbf{F_{query}}$.

$$\mathbf{F_{put}} = \mathbf{F_{ref}} - \mathbf{F_{query}} \quad (7)$$

Finally, let $\mathbf{P_{put}}$ be the table of putative functional linkages in the query genome. $\mathbf{P_{put}}$ can be obtained by performing a projection and selection operation following a cartesian product between $\mathbf{D_{q1}}$, $\mathbf{D_{q2}}$ and $\mathbf{F_{put}}$. This operation will, for each DFT in $\mathbf{F_{put}}$, enumerate all permutations of proteins that contain the first domain in the DFT to proteins that contain the second domain in the DFT. Note that this operation can be more efficiently performed if $\mathbf{F_{put}}$ includes only domains found in the query genome.

$$\mathbf{P_{put}} = \pi_{D_{q1}.seq\_id, D_{q2}.seq\_id}(\sigma_{(F_{put}.dom1=D_{q1}.dom)\wedge(F_{put}.dom2=D_{q2}.dom)}(\mathbf{F_{put}} \quad \mathbf{D_{q1}} \quad \mathbf{D_{q2}})) \; (8)$$

Remember the alternatively ordered DFTs have not been removed. Therefore, if there is a putative functional linkage between protein A and protein B, there will also be a functional linkage between protein B and protein A in $\mathbf{P_{put}}$. To remove these redundant putative functional linkages, it is easiest to re-insert the all rows in $\mathbf{P_{put}}$ into a new table with a database trigger enabled that restricts the row insertion of protein A and protein B, if the row of protein B and protein A exists.

Results and Discussion

Domain fusion analysis of H. sapiens *and* S. cerevisiae

From our domain fusion analysis on the SWISS-PROT+TrEMBL database, we identified 235 and 189 putative functionally linked protein partners in *H. sapiens* and *S. cerevisiae*, respectively (Table 1). While it is difficult to rigorously determine the accuracy of the predictions as not all protein-protein interactions have been mapped in these genomes, we searched protein interaction databases [24–26] and scientific abstracts on PubMed for those protein partners that have both gene names mentioned in the same article. The databases and scientific literature clearly indicated a functional linkage (such as gene proximity, a complex formation or common pathway) in 33 and 35 protein partners in *H. sapiens* and *S. cerevisiae*, respectively (Table 2). For example, we identified three known pairs of functionally linked proteins and one hypothetical in the *H. sapiens* and *S. cerevisiae* (Figure 2). First, a glyoxylate cycle protein [27] in *C. elegans* and *D. melanogaster* corresponds to two proteins (MASY_YEAST and ACEA_YEAST) that are known to be involved in glyoxylate cycle of *C. albicans* [28] (Figure 2a). Second, COXW_YEAST and Q12184 are involved in heme A synthesis in *S. cerevisiae* [29] (Figure 2b). Third, the levels of O76091 and FHIT_HUMAN mRNA are highly correlated in mouse homologs [30] (Figure 2c). Lastly, the functional linkage of TYSY_HUMAN and DYR_HUMAN is predicted by the domain fusions in many grain genomes including *Z. mays*, *G. max* and *A. thaliana* (Figure 2d). There was a higher percentage of positives in *S. cerevisiae* largely due to the extensive work in mapping protein-protein interactions in yeast using two-hybrid screens [31,32], microarrays [33,34] and mass spectrometry [3,4].

In *H. sapiens* sequences, there were 771 DFTs arising from 208 organisms, while in S. cerevisiae there were 1,491 DFTs from 328 organisms. The uneven sequence sampling in the SWISS-PROT+TrEMBL database skews the absolute distribution of organisms, however the distributions relative to each other are interesting. When comparing the relative changes in the distributions, the effect of uneven sampling of organisms in our database is normalized. Specifically, if the probability of finding a DFT is equal in all sequences, then the genomic distribution of the source of the DFTs would be the same as the source of the reference sequences (Figure 3a). The genomic distribution of DFTs for *S. cerevisiae* and *H. sapiens* are different to the reference sequences and to each other (Figure 3b,3c). In H. sapiens, the multicellular eukaryotic organisms (such as *M. musculus*) have advanced in the top ten sources, whereas the single cellular organisms (such as *E. coli*) have declined. Additionally, while *C. elegans* is still in the top ten sources of DFTs, it requires 476 sequences to find one DFT compared to 45 for *X. laevis* (Table 3). Conversely, in *S. cerevisiae*, single cellular organisms (such as *M. tuberculosis*) have advanced in the top ten sources and multicellular eukaryotic organisms have declined. Furthermore, distantly related organisms appear to require more sequences per DFT, yet closely related organisms do not require the fewest sequences per DFT (Table 4). One possible explanation is that domain fusions require evolutionary time to accumulate, however, if too much time passes, it may be lost. This suggests that domain fusion events cannot be used to accurately predict phylogeny, since their absence could be the result of too short or too long evolutionary distance.

Effect of our distinct definition of a fusion event

Previous methods for domain fusion analysis [13,14,20] are essentially identical to our method, except that our method specifically finds individual "domain" fusions, whereas the previous methods used full-length proteins from one organism, which correspond to a fused full-length protein in another organism. We chose our approach as many proteins consist of multiple domains. For example, consider a fusion protein in the reference organism consisting of domains ABCD, which corresponds to two separate proteins in the query organism, consisting of domains AB and CD. Using our method, the list of reference DFTs would be AB, AC, AD, BC, BD and CD; the list of query DFTs would be AB and CD. Therefore, the valid DFTs that can be used for predicting functional linkages

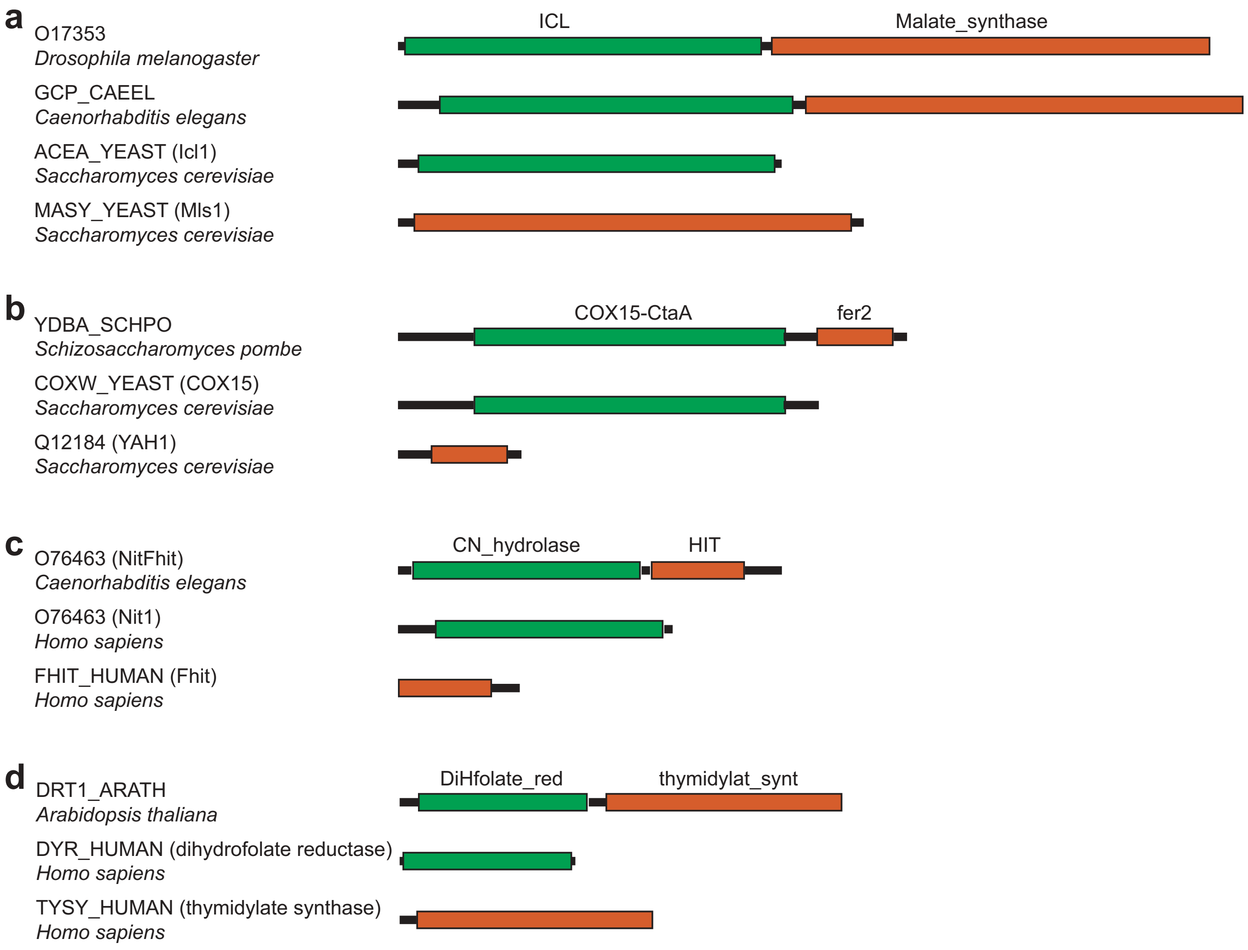

Figure 2

Examples of predicted functional linkages The sequences and domains are identified by their SWISS-PROT+TrEMBL and Pfam id, respectively, and the gene name is enclosed in brackets if applicable. The first three examples are known functional linkages in *S. cerevisiae* and *H. sapiens*, while the last one is unknown.

Table 1: Analysis of the *S. cerevisiae* and *H. sapiens* sequences in SWISS-PROT+TrEMBL

	S. cerevisiae	*H. sapiens*
Reference sequences	199,971	194,089
Query sequences	4,664	10,546
Reference DFTs	300,458	292,652
Query DFTs	13,686	67,552
Valid DFTs	290,902	237,036
Valid DFTs involving domains found in the query sequences	4,792	6,640
Putative functional linkages	415,210	4,502,378
Filtered functional linkages	189	235
Functional linkages supported by the scientific literature	35	33

Table 2: Types of functional linkages from the scientific literature

Organism	Gene proximity	Complex formation	Common pathway	Total
S. cerevisiae	1	16	18	35
H. sapiens	11	10	12	33

are AC, AD, BC and BD. All four DFTs would predict the same functional linkage between the two query organism proteins. In contrast, previous methods would have only a single fusion event that predicts this functional linkage. Therefore, an additional advantage of our approach is that the number of different DFTs predicting a functional linkage could be used to rank our prediction confidence.

Consideration of splice variants

Splice variants are treated intermediately as separate genes in our method since each variant may interact with different proteins. For example, consider a query gene with two variants: one variant consisting of domains ABC while another consisting of domains AC. If it is found that BD is a valid DFT for functional linkage prediction, then the first splice variant could be involved in a putative functional linkage that the second is not. Finally, the putative functional linkages of the gene would be the union of functional linkages of the splice variants.

Consideration of false positives

Any prediction method could produce false positive results. Here, we consider several sources of false positives, which may be generated by the present method. A false positive can occur when a functional linkage is predicted between two proteins where none exists. One possible source of false positives in domain fusion analysis is the promiscuity or paralogy in domains (for example, BTB, PDZ, SH2 and SH3 domains), which occur at a high frequency in many different protein sequences that do not share similar functions [13,14,20]. The removal of promiscuous domains reduces false positives, but the criterion for classifying them is a difficult problem. One criterion relies on finding domains with a Z-score greater than 10 [13,20], while another on domains that are involved in domain fusions events with more than 25 other domains [14].

In our analysis, instead of removing promiscuous domains altogether, we removed only promiscuous DFTs. For example, the RasGAP domain is involved in 72 functional linkages with the SH3 domain in the *H. sapiens* genes in the SWISS-PROT+TrEMBL database, however it is only involved in 2 functional linkages with the BTK domain. The DFT of RasGAP and SH3 is more promiscuous than RasGAP and BTK. Since the verification of a predicted functional linkage was performed manually through a literature search, there was a limitation to the number of linkages that could be screened. Therefore, we stringently defined a promiscuous DFT as one involved in 10 or more functional linkage predictions (Figure 4). For a higher tolerance for false positives, it is possible to use a value greater than 10.

Another possible source of false positives is the inability to list all the DFTs in the query genome. For example, consider two query genes: one consisting of a domain A while another consisting of a domain B. If it is found that AB is a valid DFT for functional linkage prediction, then the two query genes are perhaps functionally linked. However, if the query genome's DFT list is incomplete, AB may potentially exist and therefore, the two query genes may be falsely predicted as functionally linked. A number of factors can cause this problem including the use of an incomplete query genome, absent or inaccurate profile HMM domains and the erroneous prediction of intron and exon sites.

The domain fusion analysis using relational algebra presented here relies on the prediction of domains from profile HMMs. In contrast, previous approaches to domain fusion analysis often employed heuristic local pairwise sequence alignment (PSA) algorithms such as BLAST [17]. Such algorithms emphasize finding long high scoring local alignments, however, the most strongly conserved residues are commonly distributed across the domain. Therefore, the key drawback of a heuristic PSA-based approach in domain fusion analysis is its relative insensitivity for finding remote homologs and, consequently, domain fusions. Within the last decade, however, the sensitivity of sequence searching techniques has been improved by profile- or motif-based analysis, like the profile HMM, which uses information derived from multiple sequence alignments to construct and search for sequence domains and patterns [35–37]. Unlike the heuristic PSA algorithms, a profile or motif can exploit additional information, such as the position and identity of residues that are conserved throughout the domain, as well as variable insertion and deletion probabilities. Therefore, the advantage of the profile HMM is the sensitivity and accurate delineation of domains, however, the key drawback is its reliance on the accurate construction of a profile HMM for

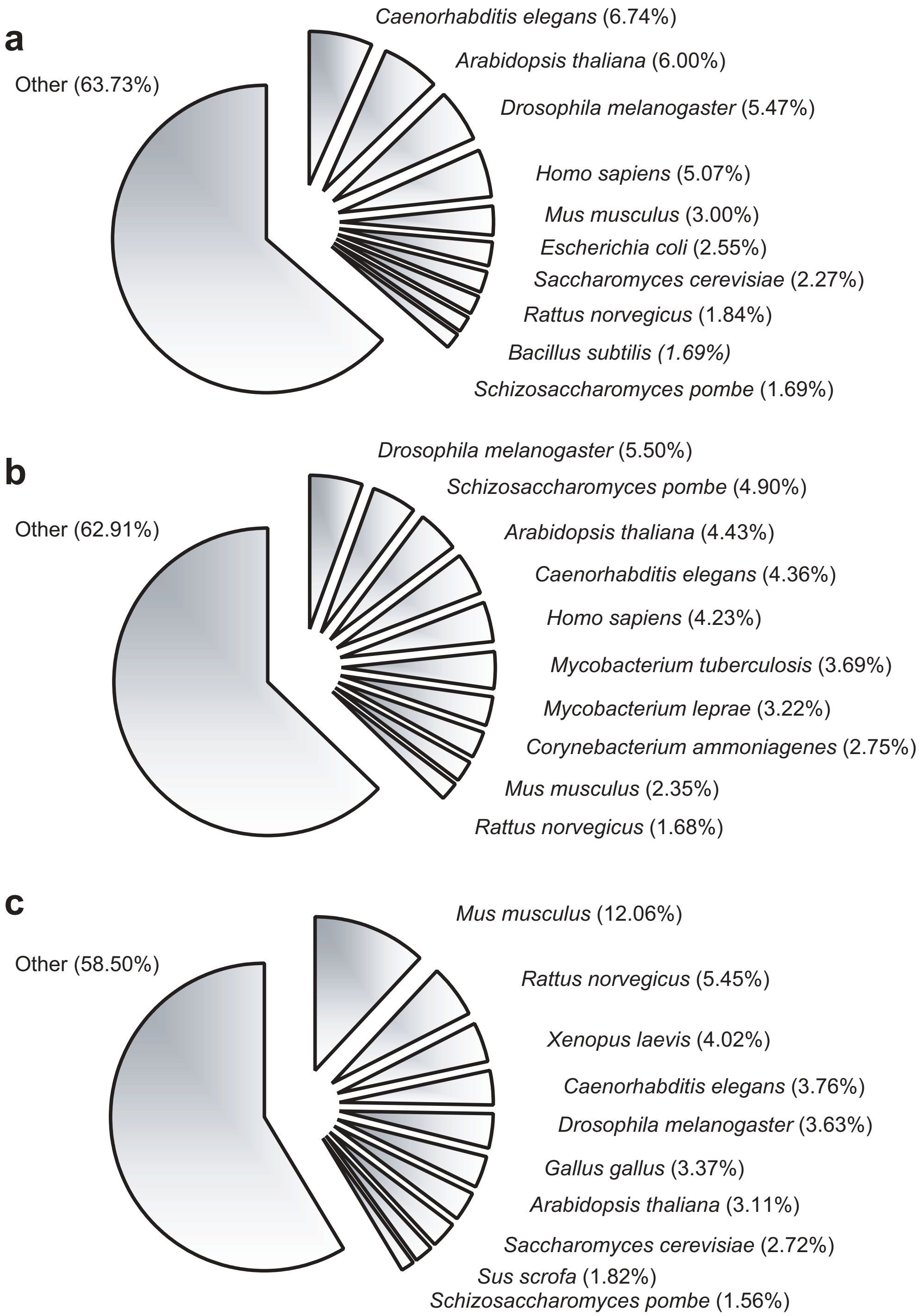

Figure 3

The genomic distribution of sequences in the (a) SWISS-PROT+TrEMBL database compared to DFTs in (b) *S. cerevisiae* and (c) *H. sapiens* The genomic distribution of the sequences in the SWISS-PROT+TrEMBL should be similar to the genomic distribution of the sources of DFTs, if the assumption is true that DFTs are equally likely to occur in all species. The figure shows clear differences in the genomic distribution, indicating that the assumption is false. Instead, the source of DFTs of a particular query genome is preferentially found in certain genomes.

Table 3: Top ten sources of DFTs in *H. sapiens*

Organism	DFTs	Total sequences in SWISS-PROT+TrEMBL	Sequences per DFT
M. musculus	93	6,146	66.1
R. norvegicus	42	3,759	89.5
X. laevis	31	1,386	44.7
C. elegans	29	13,795	475.7
D. melanogaster	28	11,187	399.5
G. gallus	26	1,398	53.8
A. thaliana	24	12,269	511.2
S. cerevisiae	21	4,653	221.6
S. scrofa	14	681	48.6
S. pombe	12	3,369	280.8

Table 4: Top ten sources of DFTs in *S. cerevisiae*

Organism	DFTs	Total sequences in SWISS-PROT+TrEMBL	Sequences per DFT
D. melanogaster	82	11,187	136.4
S. pombe	73	3,369	46.2
A. thaliana	66	12,269	185.9
C. elegans	65	13,795	212.2
H. sapiens	63	10,372	164.6
M. tuberculosis	55	3,148	57.2
M. leprae	48	1,088	22.7
C. ammoniagenes	41	37	0.9
M. musculus	35	6,146	175.6
R. norvegicus	25	3,759	150.4

all domains. If the profile HMM of a domain is not constructed or carelessly done, it will not find all putative domains and, consequently, domain fusions. Thus, as the quality and quantity of separate domain databases increases such as BLOCKS [36], PROSITE [35], Pfam [23], SMART [38], PRINTS-S [39], ProDom [40], TIGRFAMs [41] and amalgamated domain databases such as InterPro [22], our approach to domain fusion analysis will also become increasingly powerful.

Conclusions

The relational algebra method presented here offers an alternative approach to performing domain fusion analysis that leverages on existing efforts to improve the size and quality of domain and motif databases. We have illustrated the efficacy of the method by identifying many possible functional linkages in *H. sapiens* and *S. cerevisiae* sequences in the SWISS-PROT+TrEMBL database. Interestingly, the genomic distribution of the sources of DFTs suggests that DFTs are not likely found either in closely or remotely related organisms, but rather there is a balance between the two extremes that is tilted toward closely related organisms. Finally, future work could expand the method presented here to other genomes of interest.

Methods

The analysis was performed on the Oracle RDBMS (version 8) installed on a computer with a dual 750 MHz UltraSPARC-III processor and 4 G of RAM running SunOS 5.8. Sequence information from SWISS-PROT (Release 39) + TrEMBL (Release 17) and domain architecture information from Pfam was migrated to the sequence table and domain layout table of the database, respectively, by Perl and Oracle SQL*loader scripts. To perform the analysis, relational algebra expressions were converted to SQL statements and executed by an Oracle SQL*Plus client connected to the database server. The total computation time for *H. sapiens* and *S. cerevisiae* were approximately 4 and 3 hours, respectively.

Authors' contribution

KT developed the algorithm, performed the analysis and prepared the manuscript. MI provided funding and supervision for the work. Both authors have read and approved the final manuscript.

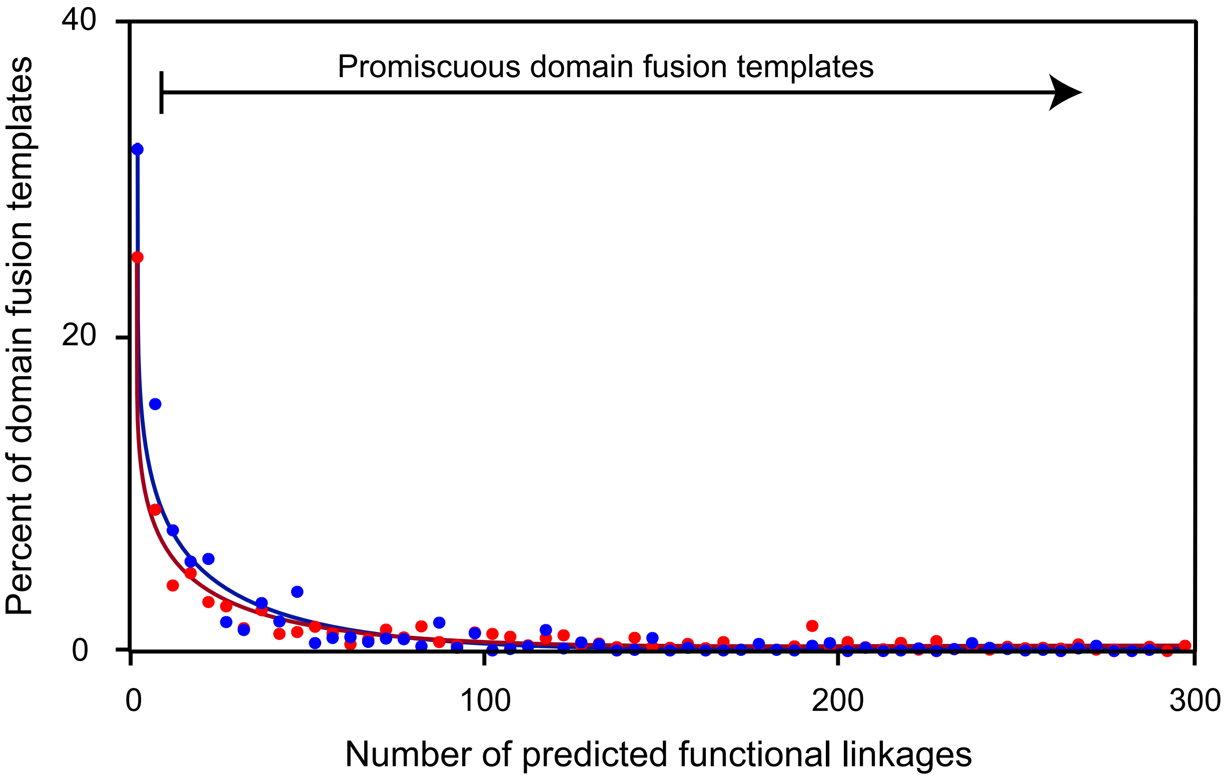

Figure 4

The percentage of DFTs to the number of functional linkages in *S. cerevisiae* (blue) and *H. sapiens* (red) As expected, the general distribution of the *S. cerevisiae* and *H. sapiens* are almost identical. Some DFTs such as ig and SH3 in *H. sapiens* and rrm and pkinase in *S. cerevisiae* predict over 5,000 functional linkages. These DFTs are promiscuous and should be removed from the analysis because they predict functional linkages between proteins that are likely unrelated. In the analysis we applied a stringent definition of a promiscuous DFT – those being involved in greater than 10 predicted functional linkages.

Acknowledgements

We would like to thank Amanda Mayo for her assistance in verifying functional linkages from protein interaction databases and scientific literature. This work was supported by a Canadian Institutes of Health Research (CIHR) fellowship to KT and a National Cancer Institute of Canada grant to MI. MI is a CIHR Investigator.

References

1. Martzen MR, McCraith SM, Spinelli SL, Torres FM, Fields S, Grayhack EJ and Phizicky EM **A biochemical genomics approach for identifying genes by the activity of their products** *Science* 1999, **286:**1153-5
2. Fields S and Song O **A novel genetic system to detect protein-protein interactions** *Nature* 1989, **340:**245-6
3. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K and Boutilier K **Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry** *Nature* 2002, **415:**180-3
4. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM and Cruciat CM **Functional organization of the yeast proteome by systematic analysis of protein complexes** *Nature* 2002, **415:**141-7
5. Jones S and Thornton JM **Principles of protein-protein interactions** *Proc Natl Acad Sci U S A* 1996, **93:**13-20
6. Larsen TA, Olson AJ and Goodsell DS **Morphology of protein-protein interfaces** *Structure* 1998, **6:**421-7
7. Overbeek R, Fonstein M, D'Souza M, Pusch GD and Maltsev N **The use of gene clusters to infer functional coupling** *Proc Natl Acad Sci U S A* 1999, **96:**2896-901
8. Dandekar T, Snel B, Huynen M and Bork P **Conservation of gene order: a fingerprint of proteins that physically interact** *Trends Biochem Sci* 1998, **23:**324-8
9. Tamames J, Casari G, Ouzounis C and Valencia A **Conserved clusters of functionally related genes in two bacterial genomes** *J Mol Evol* 1997, **44:**66-73
10. Pellegrini M, Marcotte EM, Thompson MJ, D Eisenbertg and Yeates TO **Assigning protein functions by comparative genome analysis: protein phylogenetic profiles** *Proc Natl Acad Sci U S A* 1999, **96:**4285-8
11. Marcotte EM, Xenarios I, van Der Bliek AM and Eisenberg D **Localizing proteins in the cell from their phylogenetic profiles** *Proc Natl Acad Sci U S A* 2000, **97:**12115-20
12. Deng M, Mehta S, Sun F and Chen T **Inferring domain-domain interactions from protein-protein interactions** *Genome Res* 2002, **12:**1540-8

13. Enright AJ, Iliopoulos I, Kyrpides NC and Ouzounis CA **Protein interaction maps for complete genomes based on gene fusion events** *Nature* 1999, **402:**86-90
14. Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO and Eisenberg D **Detecting protein function and protein-protein interactions from genome sequences** *Science* 1999, **285:**751-3
15. Eisenberg D, Marcotte EM, Xenarios I and Yeates TO **Protein function in the post-genomic era** *Nature* 2000, **405:**823-6
16. Huynen M, Snel B, Lathe W 3rd and Bork P **Predicting protein function by genomic context: quantitative evaluation and qualitative inferences** *Genome Res* 2000, **10:**1204-10
17. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs** *Nucleic Acids Res* 1997, **25:**3389-402
18. Tsoka S and Ouzounis CA **Prediction of protein interactions: metabolic enzymes are frequently involved in gene fusion** *Nat Genet* 2000, **26:**141-2
19. Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO and Eisenberg D **A combined algorithm for genome-wide prediction of protein function** *Nature* 1999, **402:**83-6
20. Enright AJ and Ouzounis CA **Functional associations of proteins in entire genomes by means of exhaustive detection of gene fusions** *Genome Biol* 2001, **2:**Research 0034
21. Kriventseva EV, Biswas M and Apweiler R **Clustering and analysis of protein families** *Curr Opin Struct Biol* 2001, **11:**334-9
22. Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P, Cerutti L, Corpet F and Croning MD **InterPro – an integrated documentation resource for protein families, domains and functional sites** *Bioinformatics* 2000, **16:**1145-50
23. Bateman A, Birney E, Durbin R, Eddy SR, Howe KL and Sonnhammer EL **The Pfam protein families database** *Nucleic Acids Res* 2000, **28:**263-6
24. Bader GD, Donaldson I, Wolting C, Ouellette BF, Pawson T and Hogue CW **BIND – The Biomolecular Interaction Network Database** *Nucleic Acids Res* 2001, **29:**242-5
25. Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM and Eisenberg D **DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions** *Nucleic Acids Res* 2002, **30:**303-5
26. Kanehisa M, Goto S, Kawashima S and Nakaya A **The KEGG databases at GenomeNet** *Nucleic Acids Res* 2002, **30:**42-6
27. Liu F, Thatcher JD, Barral JM and Epstein HF **Bifunctional glyoxylate cycle protein of Caenorhabditis elegans: a developmentally regulated protein of intestine and muscle** *Dev Biol* 1995, **169:**399-414
28. Lorenz MC and Fink GR **The glyoxylate cycle is required for fungal virulence** *Nature* 2001, **412:**83-6
29. Barros MH, Nobrega FG and Tzagoloff A **Mitochondrial ferredoxin is required for heme A synthesis in Saccharomyces cerevisiae** *J Biol Chem* 2002, **277:**9997-10002
30. Pekarsky Y, Campiglio M, Siprashvili Z, Druck T, Sedkov Y, Tillib S, Draganescu A, Wermuth P, Rothman JH and Huebner K **Nitrilase and Fhit homologs are encoded as fusion proteins in Drosophila melanogaster and Caenorhabditis elegans** *Proc Natl Acad Sci U S A* 1998, **95:**8744-9
31. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M and Pochart P **A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae** *Nature* 2000, **403:**623-7
32. Ito T, Tashiro K, Muta S, Ozawa R, Chiba T, Nishizawa M, Yamamoto K, Kuhara S and Sakaki Y **Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins** *Proc Natl Acad Sci U S A* 2000, **97:**1143-7
33. Lashkari DA, DeRisi JL, McCusker JH, Namath AF, Gentile C, Hwang SY, Brown PO and Davis RW **Yeast microarrays for genome wide parallel genetic and gene expression analysis** *Proc Natl Acad Sci U S A* 1997, **94:**13057-62
34. Eisen MB, Spellman PT, Brown PO and Botstein D **Cluster analysis and display of genome-wide expression patterns** *Proc Natl Acad Sci U S A* 1998, **95:**14863-8
35. Hofmann K, Bucher P, Falquet L and Bairoch A **The PROSITE database, its status in 1999** *Nucleic Acids Res* 1999, **27:**215-9
36. Henikoff S, Henikoff JG and Pietrokovski S **Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations** *Bioinformatics* 1999, **15:**471-9
37. Eddy SR **Profile hidden Markov models** *Bioinformatics* 1998, **14:**755-63
38. Schultz J, Milpetz F, Bork P and Ponting CP **SMART, a simple modular architecture research tool: identification of signaling domains** *Proc Natl Acad Sci U S A* 1998, **95:**5857-64
39. Attwood TK, Croning MD, Flower DR, Lewis AP, Mabey JE, Scordis P, Selley JN and Wright W **PRINTS-S: the database formerly known as PRINTS** *Nucleic Acids Res* 2000, **28:**225-7
40. Corpet F, Servant F, Gouzy J and Kahn D **ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons** *Nucleic Acids Res* 2000, **28:**267-9
41. Haft DH, Selengut JD and White O **The TIGRFAMs database of protein families** *Nucleic Acids Res* 2003, **31:**371-3