Joint learning of logic relationships for studying protein function using phylogenetic profiles and the Rosetta Stone method

Xin Zhang[1], Seungchan Kim[1,2], Tie Wang[1] and Chitta Baral[1]
[1]Department of Computer Science and Engineering, Arizona State University, Tempe, AZ 85287, USA
[2]Translational Genomics Research Institute, 445 N. Fifth Street, Phoenix, AZ 85004, USA

***Abstract*— Identifying logic relationships between proteins is essential for understanding their function within cells. Previous studies have been done to infer protein logic relationships using *pairwise* and *triplet* logic analysis on phylogenetic profiles. Other computational methods have also been developed using pairwise analysis on Rosetta Stone data to infer protein functional linkages[1]. In this work, we describe a Bayesian modeling framework for combining phylogenetic profile data via a likelihood with Rosetta Stone data via a prior. Based on the proposed framework, we develop a general method for jointly learning high order logic relationships among proteins whose presence or absence can be identified by logic functions[2]. The method is applied to analyze protein *triplets* and *quartets* on phylogenetic profile and Rosetta Stone datasets with 140 clusters of orthologous genes (COGs). The biological meaning of the top 30 significant triplets are further verified using the KEGG and NCBI databases. Over 50% of the discovered relationships that are associated with high significant scores could not be inferred using phylogenetic profile or Rosetta Stone data alone. Our statistical analysis shows that all significant quartets have p-values $\leq$ 5.71E-04. Many of them assign putative functional roles on uncharacterized proteins.**

***Index Terms*— phylogenetic profiles, Rosetta Stone method, protein logic relationships.**

I. Introduction

Identifying protein functions is essential for understanding molecular interactions. The functions of many translated proteins can be predicted by homology. However, homology based methods can only provide a partial understanding of a protein's function [5]. Alternatively, studying protein interacting partners can provide a more complete understanding of protein function. Various non-homology-based methods, such as phylogenetic profiles [15] and gene fusion events (also known as the Rosetta Stone method) [7][14], can be used to discover functional linkages by pairwise analysis of non-homologous proteins that are co-evolved. The protein phylogenetic profile is a numerical string of length N consisting of 0s and 1s, which represents the presence or absence of the protein in each of N sequenced genomes. Pellegrini et al [15][16] have shown that proteins with similar profiles tend to be functionally linked. The Rosetta Stone method infers protein interactions using the observation that two interacting proteins expressed separately are sometimes fused into a single chain in the same or another organism [7][14].

Various research has been done to study functional relationships among proteins using different types of data. Since each type of data sources provides only partial information, assuming that noise and bias of various data sources are largely independent, jointly learning from multiple data sources can result in more and better knowledge [3]. Bowers et al [5] studied protein relationships with pairwise analysis by choosing an optimal confidence value from four types of data. They also applied protein triplet analysis with logic functions [4] using phylogenetic profile data alone. However, no previous research has been done on jointly learning protein logic relationships.

In this paper, we present a Bayesian modeling framework that combines protein phylogenetic profile data and Rosetta Stone data to study the logic relationships among protein triplets and quartets. By incorporating functional linkage information from Rosetta Stone data, we can more reliably infer high order logic relationships among proteins. Those logic relationships can further aid in assigning biological function to uncharacterized proteins.

The rest of the paper is organized as follows. In Section II, we review related work. In Section III, we develop a Bayesian modeling framework by combining phylogenetic profile data via a likelihood with Rosetta Stone data via a prior. In Section IV, we apply the proposed framework to find protein logic relationships using public phylogenetic profile [4] and Rosetta Stone data [5] with 140 distinct families known as clusters of orthologous genes (COGs). We then discuss our results. In Section V, we summarize the proposed method and discuss future work.

II. Related Work

The phylogenetic profile method for inferring protein functional relationships is based on the assumption that functionally linked proteins are under strong selective pressure to coevolve across species. The pattern that describes the presence or absence of a protein in organisms can be obtained by searching its homologs across N organisms [4]. The sequence of a protein is compared with sequences from reference organisms using BLASTP [2]. If the BLAST E-value[3] is below a certain

[1]Proteins that share the same metabolic pathway or a common structural complex are said to be functionally linked [15].

[2]For example, using quartet logic analysis, we may discover that protein d may be present in a genome only if at least one protein a, b or c is present. We represent this logic function as $d = f(a, b, c) = a \vee b \vee c$.

[3]An E-value is the probability that, by chance, there is another alignment with a similarity score greater than the given sequence.

threshold, we believe that the protein is present (denoted by 1) in the reference organism; otherwise, it is absent (denoted by 0). A string of length N consisting of 0s and 1s is called the phylogenetic profile of the protein [9][10][15]. It is believed that proteins that are engaged in the same complex or common pathway are more likely to have similar phylogenetic profiles [6]. Pellegrini et al [15][16] have shown that proteins with similar profiles strongly tend to be functionally linked. Hence, the function of uncharacterized proteins can be predicted by the function of characterized proteins within the same cluster. Most previous research [12][15][21][22] has focused on finding pairwise similarity between profiles. However, the simple pairwise measurement is not adequate to describe the complexity of a cellular network, which may involve branching, parallel, and alternative pathways. Recently, Bowers et al [4] has proposed a logic analysis method to study protein triplets. The method searches all combinations of protein triplets where one protein (called the target) is regulated by two other proteins (called the predictors) with 8 types of logic functions. They have selected all protein triplets with uncertainty coefficient values above a certain threshold, and discovered 750,000 previously unknown relationships among protein families.

Other than phylogenetic profiles, the gene fusion method [14] has also been widely used to study functionally linked proteins. Previous work [14] showed that some pairs of interacting proteins that are expressed separately in one organism may fuse into a single protein chain in another organism. The protein chain is called the Rosetta Stone protein. This fusion event in complete genomes can be identified by sequence comparison. Due to the fact that certain genes are fused together with selective pressure during evolution, the analysis of gene fusion and division events, which is commonly known as the Rosetta Stone method, has been applied to identify protein functional linkages [7]. Statistical analysis of identified protein functional relationships can be used for protein annotation [11].

III. Bayesian modeling framework

In protein triplet analysis, let us assume that proteins j and k predict protein i with logic function f. We consider a single network structure S with $j \to i$ and $k \to i$ and assign to it the logic function $f(j,k)$ that predicts the presence or absence of protein i. We systematically analyze all combinations of triplets and assign a score to them based on how well the logic functions over two proteins can predict the target protein. In the analysis of q-order logic relationships ($q \geq 3$), the network structure is a directed acyclic graph (DAG) with $q-1$ edges from the predictors to the target protein. We first explain the case when $q = 3$, and later we expand it to higher orders.

We now present a Bayesian modeling framework to jointly learn protein logic relationships from two types of data: protein phylogenetic profiles (D_{pp}) and Rosetta Stone data (D_{rs}).

A. Learning logic relationships of protein triplets

Let $S^i_{j,k}$ denote a structure in which protein i is related to proteins j and k:

$$S^i_{j,k} = \{j \to i, k \to i\}.$$

The log posterior probability of $S^i_{j,k}$ given D_{pp} and D_{rs} is:

$$\begin{aligned} & \log P(S^i_{j,k} \mid D_{pp}, D_{rs}) \\ = \; & \log P(D_{pp}, D_{rs} \mid S^i_{j,k}) + \log P(S^i_{j,k}) - \log P(D_{pp}, D_{rs}) \\ = \; & \log P(D_{pp}, D_{rs} \mid S^i_{j,k}) + c \\ = \; & \log P(D_{rs} \mid S^i_{j,k}) + \log P(D_{pp} \mid S^i_{j,k}) + c \end{aligned} \quad (1)$$

where c is a constant for all protein triplets, and thus can be ignored. In the second equality, we use the fact that, since no prior knowledge is available for any preferred network structure $S^i_{j,k}$, the log prior $\log P(S^i_{j,k})$ over structure $S^i_{j,k}$ is uniform for all i, j and k. In the last equality, we also assume that D_{pp} and D_{rs} are conditionally independent given $S^i_{j,k}$.

The first term, $\log P(D_{rs} \mid S^i_{j,k})$, can now be represented as:

$$\log P(D_{rs} \mid S^i_{j,k}) = \log P(S^i_{j,k} \mid D_{rs}) + c' \quad (2)$$

where c' is a constant for all protein triplets, since $P(S^i_{j,k})$ and $P(D_{rs})$ are constant. $P(S^i_{j,k} \mid D_{rs})$ is the prior probability of the structure $j \to i \leftarrow k$ given the Rosetta Stone data. Since the Rosetta Stone method analyzes domain fusion events only on pairs of proteins, assuming that the protein functional linkages are conditionally independent given D_{rs}, the log prior probability over the triplet structure $S^i_{j,k}$ is decomposed as the summation of the pairwise log prior probabilities:

$$\log P(S^i_{j,k} \mid D_{rs}) = \log P(j \to i \mid D_{rs}) + \log P(k \to i \mid D_{rs}) \quad (3)$$

where $P(j \to i \mid D_{rs})$ is the confidence level that protein i is functionally linked with protein j identified by the Rosetta Stone method. In the absence of Rosetta Stone confidence information, we simply use the probability $P(j \to i) \equiv \beta$ as a prior. The default value of β is set to 0.3 in our study [4].

For the second term in Eq 1, we can learn how well the profiles of j and k predict the profile of i by maximizing the log likelihood $\log P(D_{pp} \mid S^i_{j,k})$. Since the likelihood $P(D_{pp} \mid S^i_{j,k})$ represents the probability that the structure $S^i_{j,k}$ explains the dataset D_{pp}, we choose a logic function f, given by $i = f(j,k)$, that minimizes the prediction error, so as to maximize the log likelihood $\log P(D_{pp} \mid S^i_{j,k})$.

Once an optimal logic function is found for each structure, $S^i_{j,k}$, the likelihood $P(D_{pp} \mid S^i_{j,k})$ is defined as:

$$P(D_{pp} \mid S^i_{j,k}) = U(i \mid f(j,k))$$

where f is an optimal logic function, and the uncertainty coefficient [20] $U(x \mid y)$ is defined in the interval [0,1] as

$$U(x \mid y) = [H(x) + H(y) - H(x,y)]/H(x) \quad (4)$$

where $H(x)$ is the entropy of a discrete variable x [18].

[4] As in [4], we use a threshold value of 0.3 for pairwise phylogenetic profile studies. To be consistent, we also use the same threshold value for the Rosetta Stone method.

Intuitively, $U(x|y)$ is the predictability of variable x on the basis of variable y. x is completely predicted by y if and only if $U(x|y) = 1$, whereas x is independent of y if and only if $U(x|y) = 0$. Therefore, $U(i \mid f(j,k))$ represents the probability that the profile of protein i is predicted by the logic function over the profiles of proteins j and k.

In the triplet examinations described in [4], the threshold of pairwise uncertainty coefficient is set to 0.3, and the threshold of logically combined profiles is set to 0.6. We use the same threshold values in the corresponding analysis. Given a protein triplet j,k and i, where j and k are the predictors and i is the target, we only consider the triplet in which the individual pairwise uncertainty coefficient is low [$U(i|j) < 0.3$ and $U(i|k) < 0.3$]. The low pairwise uncertainty coefficient indicates that no relationship can be identified between the predictors and the target via pairwise analysis.

B. Learning logic relationships of higher order tuples

It is likely that a protein is linked to a larger set of proteins with logic relationships that go beyond triplets. Our Bayesian modeling framework can be easily extended to a more general model for jointly learning protein logic relationships with more than two predictors.

Let $S_{pre} = \{p_1, \ldots, p_n\}$ be a set of n ($n \geq 2$) predictors of the target protein i. We can assign to protein i a function given by $i = f(p_1, \ldots, p_n)$. In other word, the profile of i is predicted by the logic function f over the profiles of S_{pre}. The structure by which protein i is related to S_{pre} is given by $S^i_{S_{pre}} = \{j \rightarrow i \mid j \in S_{pre}\}$. The log posterior probability of the structure $S^i_{S_{pre}}$, given two data sources D_{pp} and D_{rs}, is:

$$\begin{aligned} & \log P(S^i_{S_{pre}} \mid D_{pp}, D_{rs}) \\ = \; & \log P(D_{pp} \mid S^i_{S_{pre}}) + \log P(S^i_{S_{pre}} \mid D_{rs}) + c \end{aligned} \tag{5}$$

where c is a constant value for all protein sets S_{pre}, and can be ignored.

We represent the log prior probability over the structure with edge-wise decomposition:

$$\log P(S^i_{S_{pre}} \mid D_{rs}) = \sum_{k \in S_{pre}} \log P(k \rightarrow i \mid D_{rs}) \tag{6}$$

where $P(k \rightarrow i \mid D_{rs})$ is the confidence level that protein i is functionally linked with protein k, as identified by the Rosetta Stone method.

Note that the computational complexity of analysis is now increased due to the increased size of the protein set S_{pre}. Moreover, a large number of samples (i.e., a large number of organisms) is required to precisely estimate logic functions for higher order proteins. Therefore, the number of proteins to be co-analyzed will be relatively small in this case.

C. Maximizing the likelihood with proper functions

When finding optimal logic functions by maximizing the log likelihood $\log P(D_{pp} \mid S^i_{pre})$ among all 2^{2^n} possible logic functions with n predictors, some of them may not reflect the actual influence of the predictors. We give the following example:

Let a, b and c be three proteins, a and b be the predictors of c, and f be a logic function, where

$$c = f(a,b) = (a \wedge b) \vee (a \wedge \bar{b})$$

Note that by simplifying the logic function f, c is a function of a with $c = a$, regardless of b. In this case, b is a pseudo predictor of c, and has no effect on c.

Predictors that do not influence the logic function may bias the learning results since they have no actual links to the target protein. To determine the logic function between a target protein and its predictors, we only use logic functions in which every predictor plays a role in predicting the target. This leads to the notion of a *proper function*.

We say that $z = f(x_1, \ldots, x_n)$ is a proper function if, for $i = 1 \ldots n$, x_i influences z through function f.

With n predictors, the number of proper functions $p(n)$ is given by

$$\begin{cases} p(n) = 2^{2^n} - \sum_{i=0}^{n-1} \binom{n}{i} p(i), & \text{for } n \geq 1; \\ p(0) = 2. & \text{otherwise.} \end{cases} \tag{7}$$

When we consider logic functions, the input order of the predictors should not be taken into account. For example, let us assume that we have two functions $f_1 = \bar{a} \wedge b$ and $f_2 = a \wedge \bar{b}$, which are different logic functions with respect to the order of their inputs a and b. The meaning of these two functions imply that if one predictor is present and the other predictor is absent, the target protein is present. Therefore, they should be considered as the same function since they are equivalent to each other.

We say that two functions are structurally equivalent if they are identical functions, regardless of the order of the input nodes. The class of structurally equivalent functions contains all logic functions that are structurally equivalent to each other. In this paper, we also refer to the class of structurally equivalent functions as a *logic type*.

Let n be the number of predictors of a logic function. In triplet analysis ($n = 2$), there are 8 logic types [4] corresponding to 10 proper functions. In quartet analysis ($n = 3$), we have 68 logic types corresponding to 218 proper functions.

To learn the function $f(j,k)$ that maximizes the likelihood $P(D_{pp} \mid S^i_{j,k})$, we only consider proper functions and their corresponding logic types. We can find an optimal function that minimizes the prediction error by searching the profiles of proteins i, j and k. Given the profile of a target protein and its predictors, the time for finding an optimal function is linear with respect to the length of the profile. Note that the optimal function may not be unique. It is possible that a protein triplet or quartet may obey different logic relationships in different biological pathways. Therefore, a network structure may correspond to more than one optimal logic function. Since all of the optimal functions correspond to the same network structure with the same predictive error, a specific choice of a logic function from those does not affect the likelihood. However, the interpretation of finding, the relation among the presence and/or the absence of specific proteins, will be affected by such choice. In this paper, we only considered one of these optimal logic functions, which we think that leads to

TABLE I

DESCRIPTION OF PROTEINS *A*, *B* AND *C*

A	11001111111111111111111001110101111111111111111110011111111111011111111
B	10110100110011111110111011111111100001110111111111010111111100111000
C	11001100000011111100100111000011110110011111110011111111111100111000
A (COG0469):	Pyruvate kinase
B (COG0574):	Phosphoenolpyruvate synthase/pyruvate phosphate dikinase
C (COG1175):	ABC-type sugar transport systems, permease components

biologically plausible results. In practice, however, the expert molecular biologists should consider all such functions and choose the one that can be experimentally validated.

Let m be the total number of proteins, n the number of input variables in a logic function, and l the length of a given phylogenetic profile. For a Boolean function with n inputs, there are $\binom{m-1}{n} \times m$ possible cases to be considered. For each case, and for a bounded number of variables, the complexity of finding the optimal function is linear in the length of the profile. The computational complexity for inferring logic relationships over all proteins is $O(m^{n+1} \cdot l)$. For a Boolean function with number of predictors $\leq n$, we need $\Omega(2^n + n \log m)$ species in the phylogenetic profiles to identify the Boolean function [1]. When n is large, an overfitting problem may occur due to insufficient sample size. Considering the computational complexity and the statistical overfitting issue, we use $n = 2$ (triplet) and $n = 3$ (quartet) in our analysis.

Given the structure $S^i_{j,k}$ and the proper function f that maximizes the likelihood, we calculate $P(D_{pp} \mid S^i_{j,k})$ using the entropy and uncertainty coefficient of protein triplet profiles. We use the same method to compute $P(D_{pp} \mid S^i_{S_{pre}})$.

IV. ANALYSIS AND DISCUSSION

In this section, we integrate phylogenetic profile and Rosetta Stone data to find the logic relationships in protein triplets and quartets by using the previous Bayesian framework. We first describe those two types of data sources, and then discuss the results of our analysis.

A. Data Sources

We have obtained the phylogenetic profile data from a publicly available database [4] consisting of a set of binary-valued vectors describing the presence or absence of each protein family in 67 fully sequenced organisms. We choose 140 distinct families from the original dataset, known as clusters of orthologous genes (COGs), where each protein family is annotated by one or more of 20 functional categories [4]. A COG is a cluster that contains individual orthologous proteins or orthologous sets of paralogs from at least three lineages [19]. A set of genes in different species is orthologous if the genes have been evolved from a single ancestral gene [19]. The set of such orthologous genes is called orthologs. Genes that are related by duplication are known as paralogs [8]. Orthologs typically retain the same function during evolution, while paralogs may evolve into new functions.

The Rosetta Stone data of the *Aeropyrum pernix* species is obtained from the ProLink database [5]. The proteins of *Aeropyrum pernix* are categorized into 140 COGs. This number is feasible for computation on a single PC in terms of running time[5]. Note that our method can be easily implemented on a parallel cluster of computers to analyze protein logic relationships with large number of COGs. The protein-coding sequences of a genome are aligned using BLAST. A confidence value is then computed from the probability that two proteins may be found to be linked by chance, when the Rosetta Stone method is used [5]. The confidence value that protein j is functionally linked with protein i by the Rosetta Stone method is $P(j \rightarrow i \mid D_{rs})$.

TABLE II

PAIRWISE AND TRIPLET ANALYSIS OF PROTEINS COG0469, COG0574 AND COG1175.

Pairwise $U(x\|y)$	
$U(A\|B)$	0.13
$U(A\|C)$	0.22
Triplet $U(x\|f(y,z))$	
$U(A\|f_{opt}(B,C))$	0.42
Rosetta Stone Data	
$P(B \rightarrow A \mid D_{rs})$	0.77
$P(C \rightarrow A \mid D_{rs})$	0.30
Triplet Score value = -2.33	

B. Joint learning on triplets of proteins

We have applied our framework, using the phylogenetic profile and Rosetta Stone data, to study the logic relationships on protein triplets. We systematically analyzed all protein triplets and computed a score value for each of them using Eq.1. The score shows how well the two predictor proteins could predict absence or presence of the target protein, given the optimal prediction function. The triplets were ranked in descending order according to their corresponding scores. Among the triplets with high scores, we observed that many predictors fall into two functional categories: the Amino acid transport and metabolism category and the Coenzyme metabolism category. Together they predict the profile of a target protein from another functional category, the translation ribosomal structure and biogenesis category. We also observed that many proteins in the triplets with high scores belong to the same category. Some logic relationships involve proteins from category *S* which is annotated as unknown functional category. These estimated connections make intuitive sense and could provide key insight into the functional roles of these proteins.

[5]The running time of searching and analyzing the functional linkages with protein triplets and quartets is about 20 hours on a Pentium 4 2.53GHz PC.

TABLE III
DESCRIPTION OF PROTEINS D, E AND F

D	11011000010011011111100111000100010111111100010011111010011101001000
E	11011100111111111111011111000100000001110110011111111010010001010000
F	11011000000011011110100111111111111011001110110011111111111100111000
D (COG3842):	ABC-type spermidine/putrescine transport systems, ATPase components
E (COG1126):	ABC-type polar amino acid transport system, ATPase component
F (COG3839):	ABC-type sugar transport systems, ATPase components

Our method recovers all relationships among protein families in [4]. Moreover, our method finds a number of novel relationships. We have evaluated the discovered relationships via known annotations of linked proteins. The following examples show several previously undiscovered triplet relationships.

We have examined the profiles of 3 proteins A (COG0469), B (COG0574) and C (COG1175) which are described in Table I. The uncertainty coefficient scores are listed in Table II.

In pairwise analysis, the uncertainty coefficient scores of $U(A|B)$, $U(A|C)$ are 0.13 and 0.22, respectively. Both of them are below the threshold value 0.3. The triplet uncertainty coefficient score is 0.42 which is below threshold 0.6. The Rosetta Stone value between A and C is 0.3. Hence using phylogenetic profile data or Rosetta Stone data alone, we could not identify triplet relationships among proteins A, B and C. However based on our joint learning method, the triplet that proteins B and C predict protein A ends up with a top 2% significant score value under the logic function $\bar{a}+b$.

The previous three COGs fall in group G (the category of Carbohydrate transport and metabolism). The ancestral gene PpsA of COG0574 is type-I polyketide synthase and is highly similar to others from Mycobacterium leprae. The ancestral gene PykF of COG0469 is involved in Pyruvate metabolism, so as PpsA. The ancestral gene UgpA of COG1175 is probably the Sn-glycerol-3-phosphate transport integral membrane protein ABC transporter. This hypothesis is supported by a FASTA score [6] which infers that UgpA is likely to play a functional role in Pyruvate metabolism based on information extracted from the NCBI database. The above biological information supports the triplet relationships that we discovered by logic analysis.

We also examined the triplet of proteins D (COG3842), E (COG1126) and F (COG3839), which are described in Table III. In Table IV, we list the pairwise and ternary uncertainty coefficient scores using phylogenetic profiles and pairwise scores using the Rosetta Stone data.

Using phylogenetic profile data alone to infer logic relationships among D, E and F will end up with no findings because the pairwise and ternary coefficient scores are below threshold values. However, applying the proposed Bayesian framework using two data sources, we found that D could be predicated by E and F with a top 1% significant score using the logic function $D = E \wedge F$. The ancestral gene glnQ of E (COG1126) is involved in the glutamine transport ATP-binding biosynthesis and the ancestral gene malK of F (COG3839) is involved in the maltose/maltodextrin transport ATP-binding biosynthesis. Both of them are involved in the procaryotic pathway of ABC transporters. The ancestral gene of D (COG3842) is potA, which is in the same category of Amino acid transport and metabolism as gene glnQ. The ancestral gene potA has been also identified as playing a functional role in the prokaryotic pathway of ABC transporters in *E.coli*. The discovered protein triplet is validated by known protein function annotations. Furthermore, the above triplet can not be identified using a single data source.

[6]A FASTA score is a sequence alignment score using the FASTA program[17], which is used to measure the sequence similarities.

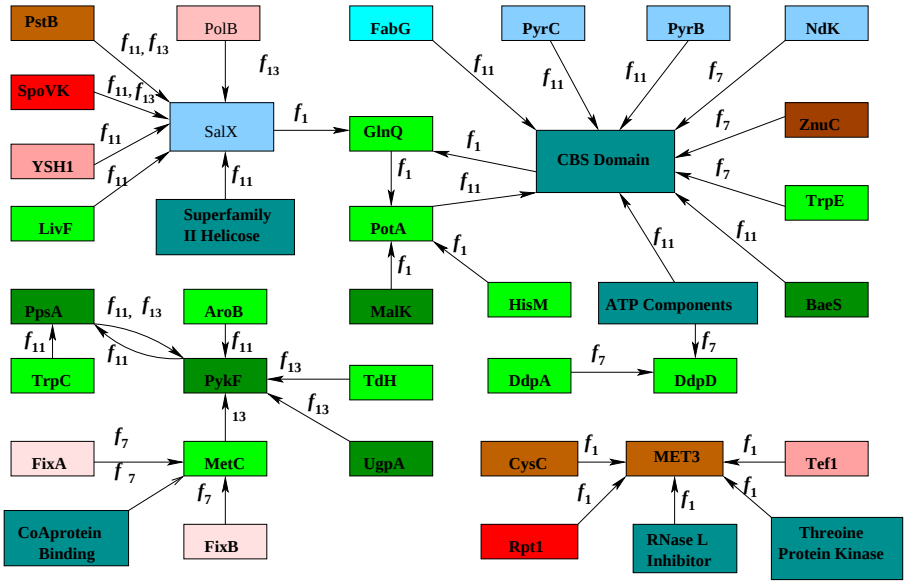
TABLE IV
PAIRWISE AND TRIPLET ANALYSIS OF PROTEINS COG3842, COG1126 AND COG3839.

Pairwise $U(x\|y)$	
$U(D\|E)$	0.18
$U(D\|F)$	0.18
Triplet $U(x\|f(y,z))$	
$U(D\|f_{opt}(E,F))$	0.48
Rosetta Stone Data	
$P(E \rightarrow D \mid D_{rs})$	0.59
$P(F \rightarrow D \mid D_{rs})$	0.60
Triplet Score value = -1.77	

The 30 most significant triplets are shown in Figure 1. The proteins in this network belong to functional categories E (Amino acid transport and metabolism), P (Inorganic ion transport and metabolism), G (Carbohydrate transport and metabolism) and C (Energy production and conversion). Among those 30 triplets, only 4 triplet relationships could be recovered using phylogenetic profiles alone and 9 functional linkages could be recovered using Rosetta Stone data alone. Combining multiple data sources can help us reveal previously unrecovered triplets. The linkages connecting uncharaterized proteins (or general function predicted proteins) with annotated proteins in the network suggest that these proteins are involved in new functions. For instance, the superfamily II Helicase, associated with COG1204, is connected to ABC-type antimicrobial peptide transport system ATPase component SalX. It suggests that the proteins in this super family might be involved in the peptide transport process.

C. Joint learning on quartets of proteins

In Eqs. 5 - 7, we showed that the proposed framework can be easily extended to infer logic relationships with more than 2 predictors. In this section, we study quartets of proteins (the target protein can be predicted by three predictors). We obtain the scores of all possible quartet combinations and rank them in descending order. The threshold score, θ, is set to -3.612 since a quartet with score -3.612 has the property



$Input_1$	$Input_2$	f_1	f_7	f_{11}	f_{13}
0	0	0	0	1	1
0	1	0	1	0	1
1	0	0	1	1	0
1	1	1	1	1	1

Fig. 1. An illustration of top 30 most significant triplets that are jointly learned by phylogenetic profile and gene fusion events. Each node represents a protein. The directed edges represent the functional linkages from the predictors to target for each triplet relationship. The edges are labeled according to the function. Nodes with the same color mean they belong to the same category, which can be obtained from supporting material from [4].

TABLE V

DESCRIPTION OF PROTEINS COG0444, COG0747, COG1136 AND COG3845

A	11111110101111111111111111000100011110001111110011111101110011011000
B	11111100101111111111111111000111100000011111110011101111110001010000
C	11111111111111111111111111111101111111111110011111111110011111111000
D	10010000000010000000100111000000011110010101110001011000011010000000
A(COG0444):	ABC-type dipeptide/oligopeptide/nickel transport system, ATPase component
B(COG0747):	ABC-type dipeptide transport system, periplasmic component
C(COG1136):	ABC-type antimicrobial peptide transport system, ATPase component
D(COG3845):	ABC-type uncharacterized transport systems, ATPase components

that three predictors can *fully* predict the target protein using phylogenetic profile data ($P(D_{pp} \mid S^{i}_{S_{pre}}) = 1.0$), while no linkage between the predictors and the target can be discovered using the Rosetta Stone method ($P(j \rightarrow i \mid D_{rs}) = 0.3, \forall j \in S_{pre}$).

The scores of protein quartets that are greater or equal to the threshold value are considered to be significant. We examined $\binom{140}{4} * 4 = 61,318,460$ combinations of all protein quartets, and discovered 143,057 previous unknown relationships with significant scores.

We examined four proteins COG0444, COG0747, COG1136 and COG3845 with ancestral genes DppD, DdpA, SalX and unknown, respectively. We list the phylogenetic profiles in Table V and pairwise, triplet and quartets analysis results in Table VI. We can identify that both DdpA and DppD play functional roles in the ABC transport pathway. The description of SalX in the KEGG database shows that the ABC-type transporter is related to transport system, which means SalX has a very close relationship with the ABC transport pathway. COG3845 is clustered with proteins whose function in the ABC-type uncharacterized transport system is general. It is highly possible that those four proteins should be annotated to be involved in the ABC transport pathway.

Note that pairwise or triplet analysis among these four pro-

TABLE VI

TRIPLET AND QUARTET ANALYSIS OF PROTEINS COG0444, COG0747, COG1136 AND COG3845.

Triplet $U(x\|f(y,z))$	
$U(A\|f(B,C))$	0.28
$U(A\|f(B,D))$	0.02
$U(A\|f(C,D))$	0.09
Quartet $U(x\|f(q,y,z))$	
$U(A\|f_{opt}(B,C,D))$	0.45
Rosetta Stone Data	
$P(B \rightarrow A \mid D_{rs})$	0.30
$P(C \rightarrow A \mid D_{rs})$	0.84
$P(D \rightarrow A \mid D_{rs})$	0.84
Quartet Score value = -2.41	

teins using phylogenetic profiles alone could not recover the linkages because they resulted in poor uncertainty coefficient scores. The relationship between DppD and DdpA is missed if we apply the Rosetta Stone value only.

The previous examples illustrate that our method can effectively reveal many quartets that are not discovered using pairwise or triplet analysis on a single data source.

In quartet analysis ($n = 3$), there are 68 logic types consisting of 218 proper functions. A total of 62 logic types were observed in quartet analysis among 140 COGs. Figure 2 shows the number of occurrences of the top 10 most frequently observed logic types based on the optimal prediction functions in quartet analysis. The corresponding logic functions are shown in Table VII. The remaining 52 logic types are not as frequently observed as those top 10 logic types. Due to space limitation, we do not list all of them in this paper.

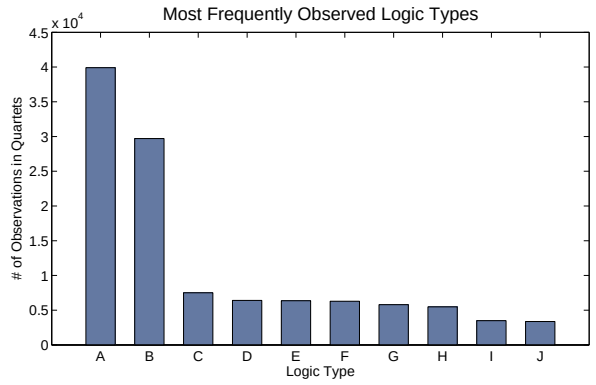


Fig. 2. Top 10 most frequently observed logic types in quartet relationships

Example 1:(Type A Logic)

COG0517 is classified only as hypothetical or putative protein in FOG:CBS domain in the NCBI database. In our analysis, COG0517 is present in a genome only if COG1121 is present, or COG3839 is not present, or COG3842 is not present. COG1121 is an ABC-type Mn/Zn transporter ATP-binding protein. COG3839 is an ABC-type sugar transporter ATP-binding protein. COG3842 is a member of the ABC-type spermidine/putrescine transport systems, ATPase components. These results suggest that the protein associated with COG0517 may play certain functional role in the ABC-type transport system as an ATP-biding protein.

Example 2:(Type B Logic)

COG1109, a phosphomannomutase, is present in a genome only if COG1208 or COG 1209 is present, or COG1319 is not present. COG1208 is Nucleoside-diphosphate-sugar pyrophosphorylase involved in lipopolysaccharide biosynthesis/translation initiation factor 2B, gamma/epsilon subunits (eIF-2Bgamma/eIF-2Bepsilon). COG1209 is dTDP-glucose pyrophosphorylase. COG1319 is Aerobic-type carbon monoxide dehydrogenase, middle subunit CoxM/CutM homologs. The results suggest that COG1109 may have putative functional linkages with pyrophosphorylase which is associated with COG1208 and COG1209, and COG1319.

The previous two examples show that the logic analysis of

TABLE VII
DESCRIPTION OF LOGIC TYPES

Logic Type	Logic Functions		
A	$a + \bar{b} + \bar{c}$	$\bar{a} + b + \bar{c}$	$\bar{a} + \bar{b} + c$
B	$a + b + \bar{c}$	$a + \bar{b} + c$	$\bar{a} + b + c$
C	$bc + a\bar{c}$	$bc + a\bar{b}$	$ac + b\bar{c}$
	$ac + \bar{a}b$	$ab + \bar{a}c$	$ab + \bar{b}c$
D	$a(b + c)$	$b(a + c)$	$c(a + b)$
E	$\bar{a} + \overline{b \oplus c}$	$\bar{b} + \overline{a \oplus c}$	$\bar{c} + \overline{a \oplus b}$
F	$a + b\bar{c}$	$a + \bar{b}c$	$b + a\bar{c}$
	$b + \bar{a}c$	$c + a\bar{b}$	$c + \bar{a}b$
G	$a + bc$	$b + ac$	$c + ab$
H	$a(b + \bar{c})$	$a(\bar{b} + c)$	$b(a + \bar{c})$
	$b(\bar{a} + c)$	$c(a + \bar{b})$	$c(\bar{a} + b)$
I	$a + \overline{b \oplus c}$	$b + \overline{a \oplus c}$	$c + \overline{a \oplus b}$
J	$a + b\bar{c}$	$b + \bar{a}c$	$c + \bar{a}b$

protein quartet can also be used to hypothesize the annotations of uncharacterized proteins or proteins that are assigned a general function.

D. Statistical Analysis

The accuracy of the discovered protein functional linkages can not be exactly verified due to a limited knowledge of protein interactions and pathways [13]. Furthermore, many protein interaction databases contain spurious linkages, which can not be directly used to evaluate our findings in terms of precision and recall.

In our work, we present statistical analysis to test the significance of discovered logic relationships. We design the method in three steps. First, we generate a matrix of randomized phylogenetic profiles maintaining the *same* individual distributions as the actual profiles. Second, we compute the score for each protein quartet using Eq.1 on the randomized datasets and rank them in descending order according to the calculated scores. Finally, we repeat the previous steps 100 times. For testing statistical significance, we used 50 nodes of dual-CPU (Xeon 2.8 GHz) machines, a subset of 512 nodes dual-CPU clusters available at ASU-TGen. It took four hours of CPU time, which results in a total computation time of 400 hours. We then evaluated the statistical significance of the discovered relationships via p-values, p_s, with respect to the log posterior probability value s, defined by [4]

$$p_s = \frac{|\mathcal{R}_s|}{|\mathcal{A}|}, \tag{8}$$

where $|\mathcal{R}_s|$ is the number of discovered logic relationships with scores $\geq s$ in the random datasets, and $|\mathcal{A}|$ is the total number of quartet trials. In this experiment, $|\mathcal{A}| = \binom{140}{4} * 100$.

We applied the method to analyze the statistical significance over 143,057 previously unknown relationships of quartets. The logic relationships discovered from the original datasets are approximately 100 times as frequent as the ones discovered from the random datasets. Figure 3 shows, in a log scale, the number of identified protein quartets against the score values for the actual and random datasets.

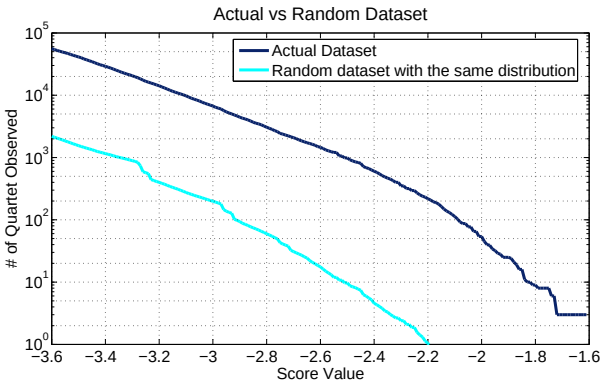


Fig. 3. A plot of the cumulative number of protein quartets recovered at a score greater than a given threshold. The number of discovered protein quartets above certain score in the actual datasets are ~ 100 times frequently as the number in the random datasets.

Statistical analysis results of the top 1,000, 2,000 and 5,000 protein quartets with significant scores s are shown in Figure 4. By analyzing the top 1,000 discovered protein quartets associated with significant scores ($s \geq -2.52$), we found that more than 76% of them have p-value $\leq$ 5E-07, and all of them have p-value $\leq$ 7E-07. We further examined 2,000 most significant quartets ($s \geq -2.69$), and observed that more than 66% of them have p-value $\leq$ 1E-06, and all of them have p-value $\leq$ 2E-06. In about 5,000 most significant quartets ($s \geq -2.93$), 96% of them have p-value $\leq$ 6.65E-06. The results also showed that all of the 143,057 protein quartets ($s \geq -3.612$) have p-value $<$ 5.71E-04.

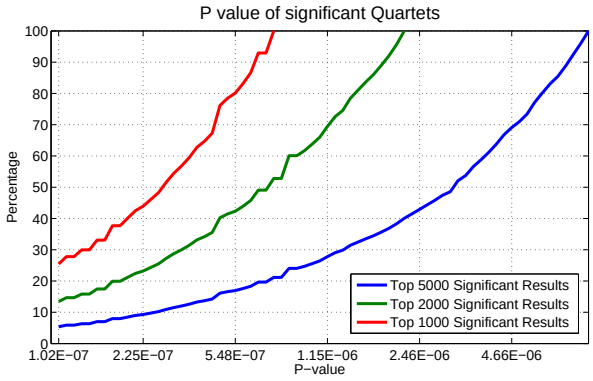


Fig. 4. p-value of the protein quartets with significant scores

V. Conclusion

In this paper, we presented a new approach for joint learning of protein logic relationships from both protein phylogenetic profile and Rosetta Stone data. We used a Bayesian model to incorporate phylogenetic profile data via a likelihood and Rosetta Stone data via a prior. By extending pairwise and triplet logic analysis, we proposed a general method for identifying high order protein logic relationships, such that the presence or absence of one protein can be predicted by the profiles of two or more other proteins. We used the notion of proper function to reflect the actual effect of the predictors on the target protein. With our generalized definitions and framework, the model can be easily extended to infer protein logic relationships with larger number of proteins.

We applied our model to jointly learn the protein triplet and quartet relationships on phylogenetic profile and Rosetta Stone datasets over 140 COGs. We identified biologically meaningful functional linkages, which could not be recovered using phylogenetic profile or Rosetta Stone data alone. In protein triplet analysis, we listed the top 30 significant protein triplets. We also applied our method to systematically examine all protein quartets. In joint learning of protein quartets, we recovered 143,057 previously unknown relationships associated with significant scores. We performed statistical analysis to evaluate the significance of the discovered protein quartets. The analysis showed that 96% of the top 5,000 quartets with significant scores have p-value $\leq$ 6.65E-06, and all of the 143,057 quartets have p-value $\leq$ 5.71E-04. A selected number of significant protein triplets and quartets were further

studied by using the KEGG and NCBI pathway databases. The putative functional linkages discovered by our joint learning method can aid in the process of annotating protein databases and help us better understand the evolution of biological systems.

VI. Acknowledgement

We thank the Guest Associate Editor Dr. John Goutsias for editing this manuscript and helping us in improving the presentation of our results. We also thank Dr. Edward B. Suh and Mr. James Lowey at TGen for helping us in running computational jobs on the ASU-TGen cluster-based supercomputer. Finally, we thank our colleague Luis Tari for contributing in the writing of the first paragraph of Section IV-A.

References

[1] Akutsu, T. and Miyano, S. Identification of genetic networks from a small number of gene expression patterns under the boolean network model. *Pacific Symposium on Biocomputing (PSB)* 1999.

[2] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, Z., Miller, W. and Lipman, D.J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 1997 Sep 1;25(17): 3389-3402.

[3] Bernard, A. and Hartemink, A.J. Informative structure priors: joint learning of dynamic regulatory networks from multiple types of data. *Pacific Symposium on Biocomputing (PSB)* 2005.

[4] Bowers, P., Cokus, S., Eisenberg, D. and Yeates, T.O. Use of Logic Relationship to Decipher Protein Network Organization. *Science*: vol 306 2246-2249 (2004).

[5] Bowers, P., Pellegrini, M., Thompson, M.J., Fierro, J., Yeates, T.O. and Eisenberg, D. Prolinks: a database of protein functional linkages derived from coevolution. *Genome Biology*: vol 5: R35 (2004).

[6] Eisenberg, D., Marcotte, E., Xenarios, I. and Yeates, T.O. Protein function in the post-genomic era. *Nature*: vol 405: 823-826 (2000).

[7] Enright, A.J., Iliopoulos, L., Kyrpides, N.C. and Ouzounis, C.A. Protein interaction maps for complete genomes based on gene fusion events. *Nature*: vol 402: (1999).

[8] Fitch, W. Distinguishing homologous from analogous proteins. *Syst Zool.*: vol 19: 2, 99-113 (1970).

[9] Gaasterland, T. and Ragan, M.A. Microbial genescapes: phyletic and functional patterns of ORF distribution among prokaryotes. *Microb. Comp. Genomics*: vol 3, 199-217 (1998).

[10] Huynen, M.A. and Bork, P. Measuring genome evolution. *Proc. Natl. Acad. Sci. USA*: vol 95 5849-5856 (1998).

[11] Krebs, W.G. and Bourne, P.E. Statistically rigorous automated protein annotion. *Bioinformatics*: vol 20 1066-1073 (2004).

[12] Liberles, D., Thoren, A., Heijne, G. and Elofsson, A. The use of phylogenetic profiles for gene prediction. *Current Genomics*: vol 3 131-137 (2002).

[13] Marcotte, E.M. Predicting protein function and networks on a genome-wide scale. *Gene Regulation and Metabolism: Post-Genomic Computational Approaches, eds. Collado-Vides, J. & Hofestadt, R.*, MIT Press (2001).

[14] Marcotte, E.M., Pellegrini, M., Ng, H.L., Rice, D.W., Yeates, T.O. and Eisenberg, D. Detecting protein function and protein-protein interactions from genome sequences. *Science*: vol 285 751-753 (1999).

[15] Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D. and Yeates, T.O. Assigning protein functions by comparative genome analysis : Protein phylogenetic profiles. *Proc. Natl. Acad. Sci. USA*: Vol 96, pp 4258-4288, 1999.

[16] Pellegrini, M., Thompson, M., Fierro, J. and Bowers, P. Computational method to assign microbial genes to pathways. *Journal of Cellular Biochemistry Supplement*: 37, pp 106-109, 2001.

[17] Pearson, W. R. Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol* 183, 63-98 (1990).

[18] Shannon, C. A mathematical theory of communication. *The Bell systems technical journal*: vol 27 379-423 (1948).

[19] Tatusov, R., Koonin, E. and Lipman, D. A Genomic Perspective on Protein Families. *Science*: vol 278: 5338, 631-7 (1997).

[20] Theil, H. Statistical decomposition analysis with applications in the social and administrative sciences. *Studies in Mathematical and Managerial Economics*: vol 14 (1972).

[21] Wu, J., Kasif, S. and DeLisi, C. Identification of functional links between genes using phylogenetic profiles. *Bioinformatics*: vol 19 1524-1530 (2003).

[22] Yanai, I. and DeLisi, C. The society of genes: networks of functional links between genes from comparative genomics. *Genome Biol.*: vol 3 0064.1 - 0064.12 (2002).

Supporting Online Material

http://www.public.asu.edu/~xzhang24/GenomicSP