AUTOMATIC MODELING OF SIGNAL PATHWAYS FROM PROTEIN-PROTEIN INTERACTION NETWORKS

XING-MING ZHAO[1,2,3], RUI-SHENG WANG[4], LUONAN CHEN[1,3,4], KAZUYUKI AIHARA[1,3]

1. *ERATO Aihara Complexity Modeling Project, JST, Tokyo 151-0064, Japan*
E-mail: xmzhao@aihara.jst.go.jp
2. *Intelligent Computing Lab, Hefei Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei, Anhui, 230031, China*
3. *Institute of Industrial Science, The University of Tokyo, Tokyo 153-8505, Japan*
4. *Department of Electrical Engineering and Electronics, Osaka Sangyo University, Osaka 574-8530, Japan*

This paper presents a novel method for recovering signaling pathways from protein-protein interaction networks automatically. Given an undirected weighted protein interaction network, finding signaling pathways is treated as searching for the optimal subnetworks from the network according to some cost function. To approach this optimum problem, an integer linear programming model is proposed in this work to model the signal pathways from the protein interaction network. The numerical results on three known yeast MAPK signal pathways demonstrate the efficiency and effectiveness of the proposed method.

1. Introduction

Signal transduction is the primary means that cells response to the external stimulus of the environment such as growth factors, nutrients, and so on. Furthermore, signal transduction plays an important role in coordinating metabolism, cell proliferation and differentiation. Generally, external signal or stimulus is transduced into a cell through an ordered sequence of biochemical reactions inside the cell. In many signal transduction processes, the number of proteins and other molecules participating in these events increases as the process proceeds from the initial stimulus, which results in a "signal cascade". Despite the success of traditional methods in detecting components involved in signaling networks, they can only generate specific linear signal pathways. The knowledge of complex signaling networks and their internal interactions is still unclear now. Therefore, it is necessary to develop new computational methods to capture the details of signaling pathways by exploiting high-throughout genomic and proteomic data.

Recently, with the advance in high-throughput bio-technology, the large-scale genomic and proteomic data provide insights into the components involved in signal transduction. For example, protein interactions and microarray data have been utilized to reconstruct signaling networks [1,2,3,4]. Since signal transduction is a process of biochemical reactions

achieved by a cascade of protein interactions, protein interaction data can provide an alternative approach to understanding signaling networks. Ideker *et al.* [4] have proposed a variant of the color coding algorithm to reconstruct signaling networks from yeast protein interaction networks. In the color coding method, a number of candidate pathways are found, with a score assigned to each candidate. The highest scoring candidate is assumed to be the putative pathway and the top scoring pathways are then assembled into a signaling network. Steffen *et al.* [2] have developed an algorithm, namely Netsearch, to reconstruct signaling networks by utilizing both gene expression data and protein interaction data. In the Netsearch method, they also rank the candidate pathways and aggregate top scoring pathways into a signaling network. Zhao *et al.* [1] have also proposed a method for ranking signal transduction pathways by utilizing both protein interaction and microarray data. In the methods described above, signaling network is not detected as a whole, on the other hand, the separate linear pathways are detected and used to assemble the signaling network.

In this work, we present a new simple and efficient method for detecting signaling pathways from protein interaction data by an integer linear programming technique. In our method, we treat the finding of signal pathways as an optimization problem and wish to find out an optimal subnetwork starting from membrane proteins and ending at transcription factors with respect to some cost functions. The objective of our method is similar to the color coding method. The difference lies in that our method treats a signaling network as a whole entity and detect it by running the model once instead of ranking individual linear pathways and assembling them into a network. The numerical experiments on yeast protein interaction data demonstrate the effectiveness of the proposed method.

The rest of the paper is organized as follows: Section 2 describes the proposed integer linear programming model; Section 3 presents the experimental results; Section 4 draws conclusions.

2. Methods

In this section, we present a method for detecting a signaling network given the possible end points (e.g. membrane proteins and transcription factors (TFs)) of signal pathways and a protein interaction network. Given a protein interaction network, it can be represented as a weighted undirected graph $G(V, E)$, where the vertices are proteins and the edge $E(i, j)$ denotes the experimentally observed interaction between proteins i and j. In this study, the weight of each edge represents the interaction reliability between the corresponding proteins. In literature, there are many methods proposed for estimating the reliability of protein interactions [5,6,7]. In this work, we utilize the method proposed by Sharan *et al.* [8] to estimate the reliability of protein interactions. With the assumption that proteins in the same signal pathway will interact with one another with high probability, the weighted protein interaction network can be utilized to find putative signaling pathways.

In the weighted network, given a starting node, the linear path of a specific length of m from the starting node to another node can be assigned a score which equals to the sum or the product of the weights for the edges in the path. With a series of paths of length m starting from specific proteins generated this way, the top ranking paths are possible

candidates for true linear signal transduction pathways. In this case, the specific starting proteins are membrane proteins because the signal transduction process starts from receptor proteins.

In this work, the weight of each edge $E(i,j)$ is defined as $a_{i,j} = -p(i,j)$, where $p(i,j)$ is the interaction reliability between proteins i and j. The score for each linear path is the sum of the weights for the edges in the path, and the length of the path is the number of proteins involved in the path. Similarly, the score of a subnetwork is the sum of the weights for the edges it contains, and the network size is the number of proteins it contains. Given an undirected weighted network $G(V, E, w)$ and the possible end points of signal pathways, i.e. membrane proteins and TFs, we wish to find out the minimum-weight subnetwork of specific size from the network G.

To accomplish the above mission, we proposed a novel integer linear programming (ILP) model to find out signal pathways, given membrane proteins, TFs and a weighted protein interaction network. The model is described as follows:

$$
\begin{aligned}
\text{Min} \quad & \sum_{i=1}^{|V|}\sum_{j=1}^{|V|} a_{ij}e_{ij} + \lambda \sum_{i=1}^{|V|}\sum_{j=1}^{|V|} e_{ij} \\
\textbf{s.t.} \quad & e_{ij} \le x_i \\
& e_{ij} \le x_j \\
& \sum_j e_{ij} \ge 1, \quad \text{if } i \text{ is a membrane protein or TF} \\
& \sum_j e_{ij} \ge 2x_i, \text{if } i \text{ is not a membrane protein or TF} \\
& x_i = 1 \text{ , if } i \text{ is a membrane protein or TF} \\
& x_i \in \{0,1\},\ i = 1, 2, \cdots, |V| \\
& e_{ij} \in \{0,1\},\ i, j = 1, 2, \cdots, |V|
\end{aligned}
$$

where a_{ij} is the weight for edge $E(i,j)$ of the undirected weighted network, x_i is a binary variable for protein i to denote whether protein i is selected as a component of the signaling network or not, e_{ij} is also a binary variable to denote whether the biochemical reaction represented by protein-protein interaction $E(i,j)$ is a part of the signaling network or not. λ is the punishment parameter to control the subnetwork size. The constraint $\sum_j e_{ij} \ge 2x_i$ is to ensure that x_i has at least two linking edges once it is selected in the signaling network so that the selected subnetwork is as connected as possible, whereas the constraint $\sum_j e_{ij} \ge 1$ makes sure that each membrane protein or TF has at least one link to other proteins. On the other hand, the constraints $e_{ij} \le x_i$ and $e_{ij} \le x_j$ ensure that only when protein i and protein j are selected as components of the signaling network, the biochemical reaction denoted by the edge e_{ij} would be considered.

The first term of the above cost function implies that we want to find out the minimum-

weight subnetwork, while the second term is used to control the subnetwork size and the number of biochemical reactions involved in the subnetwork because each protein interaction is actually a biochemical reaction. The idea behind the model is that we want to find out a minimum-weight subnetwork of specific size which accomplishes the signal transduction process with as few biochemical reactions as possible, where biochemical reactions are represented by protein interactions, i.e. e_{ij} in the cost function. The assumption is reasonable because cells usually accomplish their missions with as less energy as possibl. This criterion is also consistent with the parsimony principle widely adopted in other areas of biology such as phylogeny tree construction and haplotype inference [11,12].

The model described above is a standard integer linear programming which can be solved efficiently in polynomial time. To make the model suit for large-scale interaction networks, we can relax the constraints $x_i \in \{0,1\}, e_{ij} \in \{0,1\}$ to $0 \leq x_i \leq 1, 0 \leq e_{ij} \leq 1$ which make the ILP model become a linear programming (LP) model. Experiment results show such a relaxation does not reduce the performance, and at the same time highly improve the computation efficiency. Although the model has a parameter λ, it can be tuned in a relatively easy manner.

3. Experimental results

Our proposed ILP model was applied to find the signaling networks in the yeast protein-protein interaction network. In this work, the protein interaction data were obtained from the DIP database [9], which includes 4839 proteins and 14319 interactions. This data set has also been used by Ideker *et al.* [4] To evaluate the performance of the proposed methods, we applied it to find the three known yeast MAPK signaling pathways. The three yeast signal pathways are pheromone response, filamentous growth invasion and cell wall integrity, respectively. To reduce the computation complexity, the ILP model was applied to a smaller protein interaction network generated by depth first search (DFS) algorithm starting from membrane proteins and ending at TFs. This smaller network consists of the paths of length 6-8, and the interactions among proteins in this network were borrowed from the original protein interaction network. Therefore, three smaller protein interaction networks were generated by DFS for the three MAPK signal pathways, respectively. The sequential experiments were conducted on these three smaller protein interaction networks.

For the pheromone response pathway, the ILP model was applied to look for the signaling network starting from membrane protein STE3 and ending at transcription factor STE12. By varying the λ in the ILP model, we can get signaling networks of different size, e.g. linear pathway or signaling network. Fig.1 (a) shows the main chain of pheromone response pathway deposited in KEGG, Fig.1 (b) shows the linear signaling pathway found by color coding, and Fig.1 (c) shows the linear path found by ILP model, where the blue point is the starting point and the red one is the end point. Comparing (b) against (c), we can see that in the linear path we found, AKR1 links directly to STE5 instead of through STE4, CDC24 and BEM1 like that detected by color coding because there is a direct interaction between AKR1 and STE5. Although we failed to detect STE4, CDC24 and BEM1 in the main chain compared with color coding, we can successfully detect the linear signaling

pathway with fewer components involved in the main chain. Fewer proteins imply fewer biochemical reactions which is biologically reasonable because signals may be transduced in a parsimonious way that consume less energy.

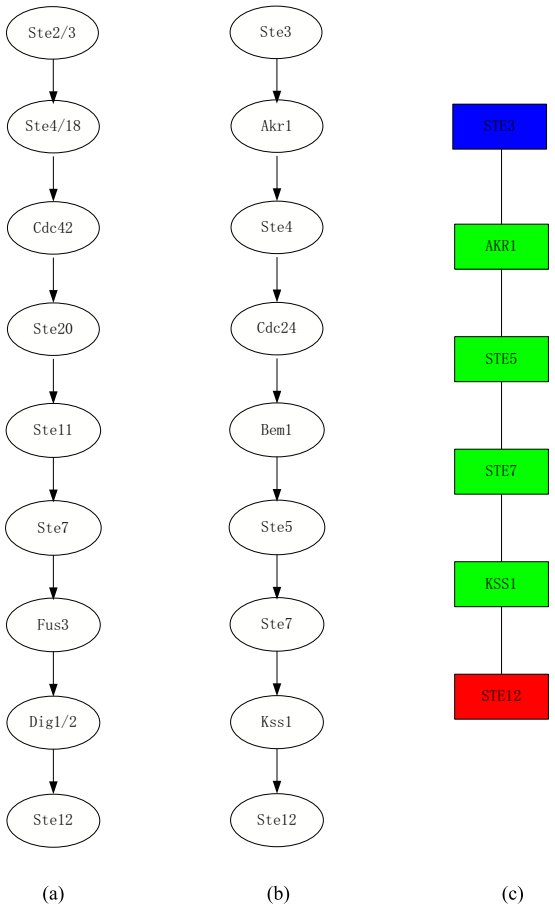

Figure 1. The linear signal pathways for pheromone response: (a) the pathway from KEGG; (b) the pathway detected by color coding; (c) the pathway detected by ILP model.

Fig.2 shows the signaling network detected by our method, where the blue point is the starting point and the red one is the end point. This signaling network consists of 19 genes. By comparing the detected signaling network with those found by Netsearch [2] and color

coding [4], we can learn that most of the components of the three signaling networks are the same. Compared with the signaling network of the same size as ours detected by Netsearch, the ILP model failed to detect proteins SST2, DIG1, DIG2 and SPH1, but detect four new proteins (STE50, BEM3, BEM4 and CDC28) which are related to the pheromone response pathway [10]. Furthermore, protein STE50 has also been detected by color coding method [4], which confirms the effectiveness of the ILP model. Compared with the color coding model, the ILP model failed to detect CDC42, DIG1 and DIG2, but detected MPT5 which has also been detected by the Netsearch method. Such a result demonstrates that our method can be a helpful complement to existing algorithms. The ILP model failed to detect DIG1 and DIG2 due to our assumption that signal transduction is assumed to be accomplished with as few biochemical reactions as possible, whereas DIG1 and DIG2 introduce many links to other proteins that have already been detected by the ILP model.

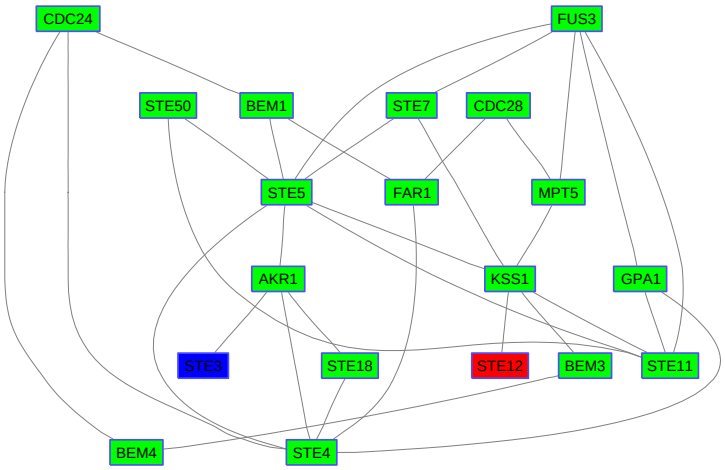

Figure 2. The signaling network for pheromone response.

For the filamentous growth invasion pathway, the ILP model was applied to detect the signaling network starting from membrane protein RAS2 and ending at transcription factor STE12. Fig.3 respectively shows the signal pathway of the same size that are deposited in KEGG, detected by color coding and ILP model, where the blue point is starting point while the red one is the end point. It can be seen from Fig.3 (a) and (c) that the signaling pathway recovered by the ILP model matches the known signal pathway to a large extent. The CDC25 and HSP82 were detected due to the missing links between RAS2 and CDC42 in the protein interaction network. Comparing Fig.3 (b) with Fig.3 (c), we can see that the ILP model can find the identical signaling pathway of the same size as that detected by color coding. Furthermore, the ILP model found out the additional links compared with the

color coding method, where the additional links may imply alternative signal pathways.

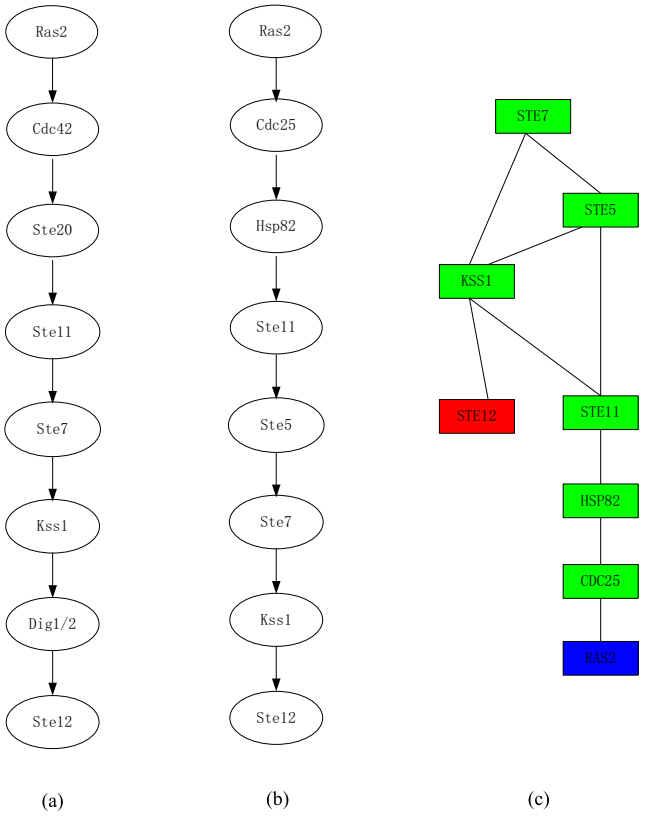

Figure 3. The signal pathways for filamentous growth invasion: (a) the pathway from KEGG; (b) the pathway by color coding; (c) the pathway by ILP model.

Furthermore, Fig.4 shows the signaling network of larger size detected by the ILP model, where the blue point is starting point while the red one is the end point. The left figure in Fig.4 shows a signaling network of size 13. Compared to the network generated by Netsearch [2], all of the proteins involved in the detected signaling network by ILP have also been found by Netsearch except GIN4, NAP1 and RIM11. The GIN4, NAP1 and RIM11

were detected because they appear in the same complex together with CDC25 [10], and GIN4 and NAP1 have the function of cell polarity and filament formation [10]. Therefore, they are related to the filamentous signaling pathway. The right figure in Fig. 4 shows another signaling network of size 19, where we assume that the proteins SPA2, CYR1, FUS3 and BEM1 are known to be involved in the signaling pathway. Although it is difficult to know exactly all the proteins involved in a signaling pathway, our assumption is reasonable because we can know some proteins in the signaling pathway from the published results by other researchers. It can be seen from Fig.4 that our detected signaling network matches that found by Netsearch [2] to a large extent. The HSC82 detected by Netsearch was not in our network because there is a direct interaction between STE11 and HSP82. The ILP model failed to detect proteins ABP1, DIG1, DIG2 and BNI1, while included two other proteins COF1 and LAS17 because COF1, LAS17, BEM1, BUD6 and SRV2 occur in the same complex [10] and therefore may have similar functions.

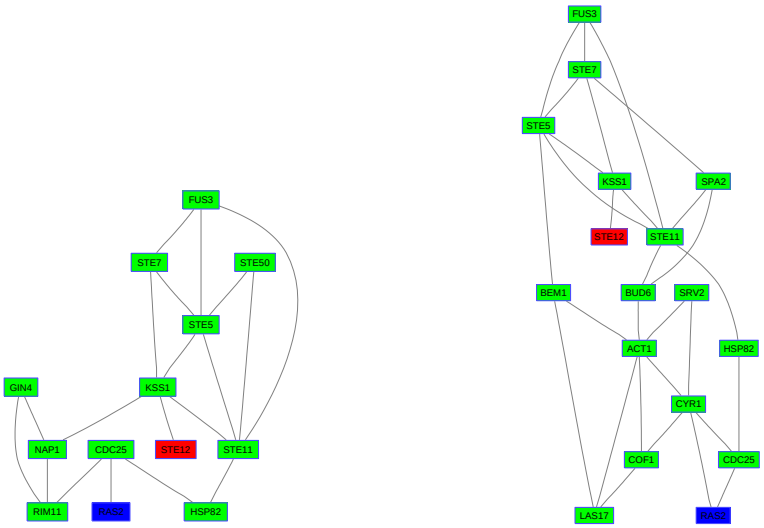

Figure 4. The signaling network for filamentous growth invasion.

For the cell wall integrity pathway, the ILP model was applied to detect the signalling network starting from MID2 and ending at RLM1. Fig.5 shows the linear signal pathways detected by the ILP model and color coding, and the one deposited in KEGG, where the blue point is starting point while the red one is the end point. It can be seen from Fig.5 that the ILP model can detect the identical signaling pathway as that by color coding. It is not surprising to see the same results because we use the same interaction data set as the one used by color coding. The detected signal pathway matches most of the known pathway except ROM2 due to the missing links between MID2 and RHO1.

From the results described above, we can see that the proposed ILP model is indeed ef-

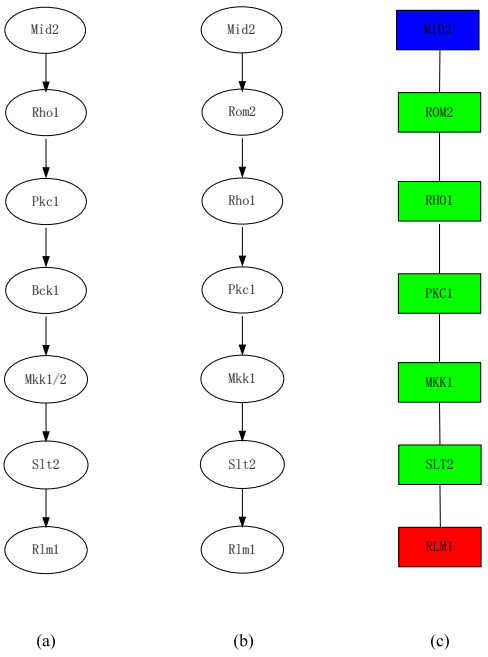

Figure 5. The linear signal pathways for cell wall integrity: (a) the pathway from KEGG; (b) the pathway by color coding; (c) the pathway by ILP model.

fective for finding signaling networks from protein interaction networks. Furthermore, the ILP model is very simple and can detect the signalling network directly instead of working in multiple-stage like Netsearch and color coding: find the candidate signal pathways, rank the candidate pathway, and assemble the top scoring pathways.

4. Conclusions

In this paper, we presented a new method for recovering signaling networks from protein interaction networks. The proposed method utilizes integer linear programming to find out the subnetwork with minimum weight of specific size. The results on three known MAPK signal pathways using yeast protein interaction network show that the ILP model can re-

cover most of the signaling pathway and the reconstructed signaling networks match most of those published results, which confirm the effectiveness and efficiency of the proposed method. Compared with existing methods, our method is much simpler because it can detect the signaling networks from protein interaction network directly instead of ranking the candidate signal pathways and assembling the top scoring signal pathways into a signaling network. Despite the success of the proposed method, it depends on the quality of the protein interactions and the estimated probabilities of the interactions. In this work, the probability of protein interactions are estimated precisely. However, most of the protein interactions are not assigned reliable scores to represent exactly the probability of protein interactions. One alternative approach to this problem is to utilize the microarray data information because there are large amount of microarray data available nowadays, and the combination of protein interactions and microarray data may provide insights into signal transduction discovery. In the future, we will explore this point in reconstructing signaling networks.

Acknowledgment

This work was partly supported by the National High Technology Research and Development Program of China (2006AA02Z309)

References

1. Liu Y, Zhao H. A computational approach for ordering signal transduction pathway components from genomics and proteomics Data. *BMC Bioinformatics*, 5: 158, 2004.
2. Steffen M, Petti A, Aach J, D'haeseleer P, Church G. Automated modelling of signal transduction networks. *BMC Bioinformatics*, 3:34, 2002.
3. Zien A, Kuffner R, Zimmer R, Lengauer T. Analysis of gene expression data with pathway scores. *Proc Int Conf Intell Syst Mol Biol*, 8:407-417, 2000.
4. Scott J, Ideker T, Karp RM, Sharan R. Efficient Algorithms for Detecting signaling pathways in Protein Interaction Networks. *Journal of Computational Biology*, 13: 133-144, 2006.
5. Bader, J., Chaudhuri, A., Rothberg, J., Chant, J. Gaining confidence in high-throughput protein interaction networks. *Nature Biotechnol*, 22: 78 - 85, 2003.
6. Deng M, Sun F, Chen T. Assessment of the reliability of protein-protein in-teractions and protein function prediction. *In: Proceedings of the Eighth Pacific Symposium on Biocomputing*, 140-51, 2003
7. von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417:399-403, 2002.
8. Sharan R, Suthram S, Kelley RM, Kuhn T, McCuine S, Uetz P, Sittler T, Karp RM, Ideker T. Conserved patterns of protein interaction in multiple species. *PNAS*, 102: 1974-1979, 2005.
9. Xenarios I, *et al*. DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.*, 30: 303-305, 2002
10. Mewes HW, Amid C, *et al*. MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Research*, 32: Database issue:D41-4, 2004.
11. Wang L, Xu Y. Haplotype inference by maximum parsimony. *Bioinformatics*, 19,1773-1780, 2003.
12. Tobias H, Andor L, Robert F, Helgi B Genetic algorithm for large-scale maximum parsimony phylogenetic analysis of proteins. *Biochim. Biophys. Acta*, 1725, 19-29, 2005.