MirrorTrees: The Alice's Wonderland of Proteomics

Sergio Castillo Joan Martí Adrià Pérez

Structural Bioinformatics, MSc in Bioinformatics for Health Sciences

1 Introduction

In the so-called OMICS era, with the high-throughput, the amount of data is huge. This is not an exception of proteomics data. Nowadays, exists huge databases for proteins, at distinct levels (inferred, detected, crystallised, etc), however, still there is a gap when concerning the interaction between proteins. This lack of knowledge is rised due to the fact that proteic physic interactions used to be found by experimental means (immunoprecipitations, double-hybtid, etc).

Now, there is the opportunity to infer physical interactions *in silico*, and then focus the experimental work to validation.

Exist different algorithms or procedures for infering proteic physical interaction, but in this essay we are going to cover only the MirrorTree approach, developed in early century by Dr. Alfonso Valencia and associates [4].

The MirrorTree is based on co-evolutionary methodology, as MirrorTree assumes that if physical interactions between proteins, they should co-evolve, moreover, they should share similar evolutionary history.

2 Methods

2.1 What is MirrorTree?

MirrorTree methodology, developed by Dr. Alfonso Valencia and associates, predicts direct interactions between proteins based on co-evolution of interacting proteins. [4]. The work-flow of the MirrorTree is as follows:

1. Gather n sequences to test its interaction.
2. Create n sets, consisted of orthologous sequences(search by homology, using BLAST [1]).
3. Build a MSA for n sets.
4. Build distance matrix for each set (using McLachlan 71).
5. Compare the the distance matrices at 1:1 rate.
6. Build a correlation coefficient (Pearson) based on the difference between the matrices.
7. Infer interaction.

MirrorTree is settled upon the following suppositions:

1. Proteins that interact directly co-evolve.
2. Proteins that interact directly and co-evolve should share the same evolutionary history, so they must have similar phylogenies.

In further sections there would be a discussion over these assumptions.

2.2 Pipeline

Our integrative tool takes a slightly different approach to the Valencia's team to MirrorTrees. Our pipeline is as follows:

1. Gather n sequences to test its interaction.
2. Create n sets, consisted of orthologous sequences(search with a hidden markov model, using jackhmmer [3]).
3. Build a MSA for n sets (using hmmalign [3]).
4. Use the distance matrix for each set provided by jackhmmer [3] (using Blosom62 matrix).
5. Compare the the distance matrices at 1:1 rate.
6. Partial out the spurious correlation (species-lineage relationship) using the distance matrix of the species tree (18S rRNA) [6].
7. Build a correlation coefficient (Pearson's r and Spearman's rho) based on the difference between the matrices.

2.3 Databases

In order to develop the tool, different train/test sets where developed. Here are listed the databases from where we have retrieved the IDs or sequences:

1. Sequences in FASTA format (Uniprot-Swiss database).
2. ID's from proteins that interact (IntAct database [2]).
3. ID's from proteins that do not interact (Negatome database [7]).
4. Alignments of 18S rRNA for the species tree (SILVA format [5]).

3 Discussion

Despite being a possible theoretical approach, Valencia's work in 2001[4] needs to be improved and analysed. The approach here presented has prominent differences with the reference work. In the following sections there are specified which points needed to be changed and the reasons for doing so.

3.1 Assumptions

One of the main handicaps that exist in MirrorTree approaching is that its assumptions are not always valid.

1. *Proteins that interact directly co-evolve*: Not necessarily. Interaction between proteins occurs at domain level, more precisely it interacts with the interaction region, a smaller region within the domain, not at the whole protein. It might be true that interaction regions co-evolve, but that does not mean that co-evolve the whole protein. Besides, even the co-evolution between interaction sites is somehow obscure, as it depends on the intensity of the selective pressure between them. It is known that the interaction between proteins is not static, is partially adaptative (key-lock vs greeting-hands). So, would always be relevant an aminoacid exchange? How intense is the selective pressure between the regions?
2. *Proteins that interact directly and co-evolve should share the same evolutionary history, so they must have similar phylogenies.*

3.2 Search of orthologs

One of the main difficulties when working with phylogenies, and by extension to this essay, is the difficulty to distinguish orthologs from paralogs, one of the paradigmatic problems on the filed.

When retrieving orthologous sequences, paralogous sequences are also retrieved. As they refer to different events of diversification (species vs. gene respectively), there is an introgression of an other relationship within our model, adding confounding relationships, and therefore decreasing the predictive power of the approach.

Besides, proteins, and in special, proteins from multicellular eukaryots, may have many domains. If the aim of the study is to asses direct interaction between proteins, it must focus on the interaction region. Using BLAST-based searches can also lead to misleading homology, as it might search homologous sequences based on a domain which is not related to the interaction.

On the contrary, HMM-based searches look for homologous sequences using an specific domain, which is fetched from the query sequence. Doing so, relationship between proteins, focusing on the interacting domain, can be ensured.

3.3 Weight Matrix

The distance matrix can be computed using different weight matrices (p.e. Blosum62, McLachlan71, Dayhoff Pam Matrix, etc) or distance correctors (p.e. Kimura's distance). Despite being a variable in the approach, some authors[8] have found that its change does not provide an improvement of the prediction power.

This might be true due to that the focus of MirrorTree is to compare between trees (a.k.a distance matrix). The relevant point is to use the same statistic in both MSAs, not in how good or accurate is the estimation of the distances.

3.4 Correlation measures

In Valencia's article[4] Pearson's r is used to test the relationship between matrices (which can be interpreted as two-dimensional ordered lists of data). Pearson's r holds its purpose upon two main suppositions:

1. *It is based on normality, or at least performs way better under normal distribution of the data:* It is not clear that the distances between genes, given a clade follow a normal distribution, so it might be an infraestimation of the intensity of the relationship.

2. *It only captures linear relationships:* Exists other kind of relationships: exponential, logarithmic, parabolic, etc. Is the relationship between 2 tree distances always linear?

3. *It is sensible to outlayers:* As it tries to fit a line in a cloud of points, the outlayers play a significant role when fitting it, biasing the fitting towards them. The fitting line would not be accurate.

As there is uncertainty in these 3 elements, it would be interesting to complement the prediction with Spearman's rho, which is rank-ordered-based (non-parametric), can fit a line or a curve in logarithmic and exponential relationships (but not parabolic, though) and it is more robust to outlayers.

In addition, correlation measures can be more reliable if species-lineage relationship is partialled out[6]. This species-lineage free relationship can be obtained through correlating the error components obtained from the correlations of a matrix distance from a query sequence and the matrix distance obtained with species-lineage marker, such as 18S rRNA [6].

4 Results

References

[1] Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403 – 410, 1990.

[2] Henning Hermjakob, Luisa Montecchi-Palazzi, Chris Lewington, Sugath Mudali, Samuel Kerrien, Sandra E. Orchard, Martin Vingron, Bernd Roechert, Peter Roepstorff, Alfonso Valencia, Hanah Margalit, John Armstrong, Amos Bairoch, Gianni Cesareni, David James Sherman, and Rolf Apweiler. Intact: an open source molecular interaction database. *Nucleic Acids Research*, 32(Database-Issue):452–455, 2004.

[3] L. Steven Johnson, Sean R. Eddy, and Elon Portugaly. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC bioinformatics*, 11(1):431+, 2010.

[4] Florencio Pazos and Alfonso Valencia. Similarity of phylogenetic trees as indicator of protein–protein interaction. *Protein Engineering*, 14(9):609–614, 2001.

[5] Christian Quast, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. The silva ribosomal rna gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, 41(D1):D590–D596, 2013.

[6] Tetsuya Sato, Yoshihiro Yamanishi, Minoru Kanehisa, and Hiroyuki Toh. The inference of protein–protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships. *Bioinformatics*, 21(17):3482–3489, 2005.

[7] Pawel Smialowski, Philipp Pagel, Philip Wong, Barbara Brauner, Irmtraud Dunger, Gisela Fobo, Goar Frishman, Corinna Montrone, Thomas Rattei, Dmitrij Frishman, et al. The negatome database: a reference set of non-interacting protein pairs. *Nucleic acids research*, 38(suppl 1):D540–D544, 2010.

[8] Hua Zhou and Eric Jakobsson. Predicting protein-protein interaction by the mirrortree method: Possibilities and limitations. *PLoS ONE*, 8(12):1–9, 12 2013.