UNIVERSIDAD AUTÓNOMA DE MADRID

Improving Co-evolution Based Methods for Protein-Protein Interaction Prediction

Author:
David Ochoa

Supervisor:
Dr. Florencio Pazos

A thesis submitted in partial fulfillment for the
degree of Doctor of Philosophy

in the
Facultad de Ciencias
Departamento de Biología Molecular

Madrid, May 2013

Declaration of Authorship

I, David Ochoa, declare that this thesis titled, "Improving Co-evolution Based Methods for Protein-Protein Interaction Prediction" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.

Signed: _____

Date: _____

Contents

UNIVERSIDAD AUTÓNOMA DE MADRID

Abstract

Facultad de Ciencias
Departamento de Biología Molecular

Improving Co-evolution Based Methods for Protein-Protein Interaction Prediction

by David Ochoa

The study of protein-protein interactions and how these interactions determine the function and behavior of the living systems is nowadays one of the fundamental questions of Systems Biology. The emergence of a number of experimental techniques providing protein interaction data at a genome scale have boosted the study of biological problems that can be studied now considering the complete network of interactions not just as the sum of the parts. Nevertheless, these experimental procedures usually suffer from technical problems, producing poor coverages or large numbers of false positives.

As an alternative to complement the experimental methods, a set of computational approaches have tried to take advantage of the different evolutive landmarks that interacting proteins print on their genes. For instance, evidence suggests that functionally related and potentially interacting proteins tend to evolve in a coordinated way, thereby presenting similar phylogenetic trees. A particularly successful family of methods, known as *mirrortree*, has emerged to quantify this co-evolution at a sequence level as a sign of interaction at a molecular level. Over the last decade, this family of techniques has demonstrated its ability to perform genome-wide protein interaction predictions, even reaching accuracies similar to their experimental counterparts. However, a number of problems affecting protein interaction prediction have appeared, either derived from technical issues or inherited from the incomplete understanding of the underlying evolutionary process. Over the coming years, these problems may have a dramatic impact on the global performance of the methods.

The main proposal of this thesis is to diagnose the aforementioned problems limiting the full implementation of co-evolution-based prediction of protein interactions, in order to offer possible solutions, potential applications and foreseeable developments. During the last years, the *mirrortree*-based family of techniques has largely been used to predict protein interactions mostly in genome-wide computational experiments. Nevertheless, the co-evolution-based prediction has also shown adequate when single pairs of proteins need to be evaluated. As a consequence, we present the Mirrortree Server, which allows users with any level of expertise to graphically and interactively study the co-evolution of two protein families in a taxonomic context. More difficulties arise when the co-evolution analysis is performed for large sets of potentially interacting proteins. Since little is known about the latent evolutionary signal, whether the tree similarity is due to compensatory changes or to similarities in the evolutionary rate, is a pivotal question that will

condition future research on this issue. We evaluate the true extent of previous discussions in this regard by incorporating predicted solvent accessibility to *mirrortree*-based predictions, which also allowed to improve performance predicting some types of interactions. Other problems arise as a consequence of the growing number of sequenced organisms available. In that sense, we show that the performance of *mirrortree*-based methodologies depends on the set of organisms used to build the trees and how the selection of certain subsets of organisms seems to be more suitable for the prediction of certain types of interactions. Finally, considering all the aforementioned analysis, we propose a new *mirrortree*-based method denominated *p-mirrortree* which calculates the statistical significance of a given tree similarity based on a null distribution of random co-evolution. Moreover, important information on the structure, function, and evolution of macromolecular complexes can be inferred with this methodology.

UNIVERSIDAD AUTÓNOMA DE MADRID

Resumen

Facultad de Ciencias
Departamento de Biología Molecular

Improving Co-evolution Based Methods for Protein-Protein Interaction Prediction

por David Ochoa

El estudio de las interacciones proteína-proteína y de cómo dichas interacciones determinan la función y el comportamiento de los sistemas vivos es hoy en día una de las preguntas fundamentales de la Biología de Sistemas. La aparición de una serie de técnicas experimentales capaces de identificar interacciones a escala genómica, ha impulsado el estudio de los problemas biológicos, no sólo como la suma de las partes, sino considerando la red completa de interacciones. Sin embargo, estos métodos experimentales a menudo adolecen de una serie de problemas técnicos que derivan en bajos rendimientos y alto número de falsos positivos.

Un conjunto de métodos computacionales han surgido como alternativa a los técnicas experimentales con el objetivo de predecir interacciones basándose en los distintos tipos de marcas evolutivas que las proteínas que interaccionan dejan en el genoma. En este sentido, ciertas evidencias sugieren que proteínas funcionalmente relacionadas que potencialmente podrían interaccionar tienden a evolucionar de una forma coordinada y por tanto poseen árboles filogenéticos similares. Una familia de métodos conocida como *mirrortree* ha surgido con el objetivo de cuantificar la coevolución a nivel de secuencia como un síntoma de interacción a nivel molecular. A lo largo de la última década, esta familia de técnicas ha demostrado predecir interacciones entre proteínas con una precisión similar a las técnicas experimentales. A pesar de ello, han surgido una serie de problemas relacionados, bien con inconvenientes de tipo técnico, o bien debidos al desconocimiento de los procesos evolutivos subyacentes. Durante los próximos años, estos problemas pueden tener un impacto dramático en el uso de estos métodos de predicción.

El principal objetivo de esta tesis es del diagnóstico de los problemas que dificultan la completa implantación de los métodos basados en coevolución con el objetivo de ofrecer posibles soluciones, potenciales aplicaciones y mejoras futuras. A lo largo de los últimos años, esta familia de técnicas se ha usado mayoritariamente para predecir interacciones entre proteínas en organismos completos. Sin embargo, la predicción basada en coevolución ha resultado ser útil también para predecir interacciones entre pares de proteínas aisladas. Por ello, presentamos MirrorTree Server, un servidor que permite a usuarios con distintos niveles de experiencia estudiar de manera interactiva la coevolución de un par de familias de proteínas en un contexto taxonómico. Sin embargo, cuando aplicamos los mismos conceptos a la predicción de interacciones para un organismo completo aparecen una serie de problemas que es necesario abordar. Puesto que se conoce poco acerca de la señal evolutiva responsable de dicha similitud, se ha postulado que por un lado puede deberse bien a cambios compensatorios a nivel de secuencia, o bien a cambios concertados debido

a una similitud en la tasa evolutiva de las proteínas. Con el objetivo de aclarar este dilema y de mejorar la predicción de interacciones, hemos estudiado el efecto de incorporar información de accessibilidad predicha en la predicción de interacciones basada en *mirrortree*. Por otro lado, otros problemas surgen como consecuencia del número creciente de organismos secuenciados. En este sentido, hemos querido explorar la precisión de las tecnologías basadas en *mirrortree* en función del conjunto de organismos que se utiliza para construir los árboles filogenéticos y cómo ciertos subconjuntos pueden ser más adecuados para determinados tipos de interacciones. Finalmente, teniendo en cuenta todos los problemas aquí descritos, proponemos un nuevo método basado en *mirrortree* denominado *p-mirrortree* que calcula la significancia estadística de un determinado valor de similitud basado en una distribución nula de coevoluciones aleatorias. Además, información relevante sobre la estructura, función y evolución de los complejos macromoleculares puede ser extraída de la aplicación de esta metodología.

Abbreviations

AD	**A**ctivator **D**omain
ATPase	**A**adenosine **T**ri**P**hosphate hydrol**ase**
AUC	**A**rea **U**nder ROC **C**urve
BBH	**B**est **B**idirectional **H**it
BD	**B**inding **D**omain
CBP	**C**almodulin **B**inding **P**rotein
cDNA	**c**omplementary **D**eoxyribo**N**ucleic **A**cid
CM	**C**ontext**M**irror
DNA	**D**eoxyribo**N**ucleic **A**cid
EGTA	**E**thylene **G**lycol **T**etraacetic **A**cid
FN	**F**alse **N**egatives
FP	**F**alse **P**ositives
FPR	**F**alse **P**ositive **R**ate
FRET	**F**luorescence **R**esonance **E**nergy **T**ransfer
HCA	**H**ierarchical **C**o-evolutionary **A**nalysis
HGT	**H**orizontal **G**ene **T**ransfer
IgG	**I**mmuno**g**lobulin **G**
ITC	**I**sothermal **T**itration **C**alorimetry
MS	**M**ass **S**pectrometry
MSA	**M**ultiple **S**equence **A**nalysis
MT	**M**irror**T**ree
N	**N**egatives
NADH	**N**icotinamide **A**denine **D**inucleotide
NMR	**N**uclear **M**agnetic **R**esonance
P	**P**ositives
PC	**P**rofile **C**orrelation
pMT	**p**-**M**irror**T**ree
PP	**P**hylogenetic **P**rofiles
PPI	**P**rotein **P**rotein **I**nteraction
PPV	**P**ositive **P**redictive **V**alue
RNA	**R**ibo**N**ucleic **A**cid
RNase	**R**ibo**N**ucle**ase**

ROC	**R**eceiver **O**perating **C**haracteristic
rRNA	**r**ibosomic **R**ibo**N**ucleic **A**cid
SPR	**S**urface **P**lasmon **R**esonance
TAP	**T**andem **A**ffinity **P**urification
TEV	**T**obacco **E**tch **V**irus
TN	**T**rue **N**egatives
TP	**T**rue **P**ositives
TPR	**T**rue **P**ositive **R**ate
UI	**U**ser **I**nterface
Y2H	**Y**east two(**2**) **H**ybrid

". . . the whole organisation is so tied together during its growth and development, that when slight variations in any one part occur, and are accumulated through natural selection, other parts become modified. This is a very important subject, most imperfectly understood."

Charles Darwin, *On the origin of Species*, 1859

Introduction

Most biological processes can hardly be understood without a comprehensive analysis of a large number of molecular components and interactions. From the simplest systems to complex mammal machineries, the interactions between different molecules usually determines the resulting phenotype. This is the case with cellular proteins, which rarely work in isolation but are frequently involved in pathways and interaction networks. The eventual perturbation of these networks can lead to disease or even death. So, an elucidation of protein-protein interactions can greatly contribute to our understanding of living systems in general and pathological states in particular.

1.1 Systems Biology and Biological Networks

Systems Biology relies on the study of the aforementioned complex networks of interactions within biological systems [1, 2]. Contrary to more traditional reductionism, which tries to understand biological problems dividing the systems on their constituent parts, Systems Biology approaches these problems from a holistic point of view. Indeed, reductionism has been extremely successful explaining many areas of Biology but, for many reasons, the integration of data is necessary to get insight into systems that cannot be understood as the sum of the parts. This change in the paradigm, together with the massive amount of data produced as consequence of what is known as the -Omics era, supposed a revolution in the study of biological systems, especially during the last 15 years. As a result, new methodologies and infrastructures previously applied to other scientific areas are now applied to solve biomedical problems. One of these new areas is the study of biological networks.

1.1.1 Biological Networks

The study of interactions between the components of biological systems and how these interactions give rise to their function and behavior can be modelled using network concepts [3]. Many phenomena can be modelled as networks, in which the entities are represented as nodes and their relationships as the connecting edges. This approach previously applied to social networks [4–6], food webs [7] or the World Wide Web [8] brings another analytical view of the complex relationships at a molecular level.

In the same way the structure of the power supply network affects the robustness and stability of power transmission, the architecture of biological networks informs about many characteristics of the systems they represent. In that sense, the topology of real networks has some important characteristics:

- *Scale free*: most of the nodes only have a few interactions, and these coexist with a few highly connected nodes, the hubs, that hold the whole network together. The number of incident edges in a given node is known as the degree. Indeed, the number of nodes with a given degree follows a power law distribution [9]. This property, sometimes referred as *scale free*, has been reported in organisms for which protein-protein interaction and metabolic networks have been determined, from yeast to human. It has been suggested that this architecture provide robustness and error-tolerance to the network since random perturbations are less deleterious than in random networks. [10].

- *Small world*: any random pair of nodes can be easily connected with a path of a relatively small number of links, independently of the size/complexity of the network [11]. For example, the metabolic network of a simple parasitic bacterium has the same average path length as the highly developed network of a large multi-cellular organism [12]. That property indicates that any local perturbation can reach the whole network very quickly.

- *Transitivity*: this graph theory concept is fulfilled when the network is much more highly clustered than a random graph, in the sense that if A is linked to B and B is linked to C, there is a greatly increased probability that A will also be linked to C. Therefore, biological networks are usually enriched in subgraphs of highly inter-connected groups frequently involved in a common function [13].

It is impossible to ignore the apparent universality of topological characteristics in real networks, including biological ones. Whether these concepts should be used to explore the dynamics, evolution or robustness is still an area of research, but many successful applications of network concepts to biological systems have been published during the last few years. However, the relationships between the molecular entities are often hard to describe. Interactions can be conditional or contextual, and may not always be captured in a given study, regardless of its attention to quality. The modelling of accurate networks remains challenging in different areas such as protein-protein interaction networks.

1.2 Protein-Protein Interaction Networks

One of the most studied biological networks are those representing Protein-Protein Interactions (PPIs). PPIs are the result of a physical binding between two or more proteins to accomplish a biological function. They can be classified into different categories depending on their strength (permanent and transient), specificity (specific or nonspecific) or the similarity between interacting

subunits (homo- and hetero-oligomers). In the next sections, we will attend to another classification depending on the evidence reporting the interaction. We will distinguish between those cases in which the protein interaction is defined based on direct or indirect physical evidence, from those based on broader functional relationships.

Protein interactions are key in many cellular processes related with functions as diverse as signalling, transport or catalysis. The perturbation of these interactions, in particular the distortion of protein interface, usually implies the development of diseases and thus the great interest in identifying protein interactions.

1.3 Experimental Determination of PPIs

The instability and heterogeneity of PPIs make their detection extremely sensible to the experimental setup. Therefore, a number of methods have been developed to screen interactions using different approaches at both, small and large scale.

Traditionally, evidence is gathered using small scale experiments designed to identify and validate a small number of targeted interactions [14]. Information defining protein interfaces at an atomic level can be provided by X-ray crystallography and NMR spectroscopy. However, the number of solved protein complexes is still low [15]. Alternatively, several spectroscopic techniques characterize protein interactions in real-time using different labels [16, 17]. For example, Fluorescence Resonance Energy Transfer (FRET) takes advantage of the modification in the spectral activity occurred when 2 fluorophores are close to each other [18]. Another method, known as Surface Plasmon Resonance (SPR), has proven effective without spectroscopic labeling in detecting interactions between soluble ligands and immobilized receptors [19, 20]. Isothermal Titration Calorimetry (ITC) allows to measure the enthalpy of binding [21]. Other methods, such as atomic force microscopy [22], can accurately analyze single molecules measuring the microscopic forces that bind the interacting proteins, or detecting conformational changes using fluorescence [23].

Nevertheless, during the last few years a set of high-throughput methodologies has emerged to exhaustively probe all the potential interactions within entire genomes. We will review some of the most used approaches, in order to understand their benefits and limitations.

1.3.1 Yeast two-hybrid

One of the most successful techniques that has accelerated the screening of interactions *in vivo* is the yeast two-hybrid (Y2H) methodology. Y2H takes advantage of the fact that eukaryotic transcription activators have at least two different domains, one that binds to the DNA promoter (BD) and another that activates the transcription (AD). It has been reported how the transcription can be disrupted by splitting the activator in their two constituent domains, whereas the activity can be restored if BD is physically associated with AD [24]. This modular property allows both domains to function in proximity to each other without direct binding. Under this principle, plasmids are engineered producing protein products containing BD and AD fused to the 2 proteins whose

interaction is being assayed. The chimeric sequences are usually referred to as bait and prey respectively. If the two proteins interact, a downstream reporter gene is activated, producing a detectable phenotype. Although the typical setup often involves beta-galactosidase as reporter in yeast, numerous variations of Y2H have been developed including: systems with several reporter genes, one and three hybrids to identify interactions with DNA and RNA [25–28], detection of interactions in mammalian and prokaryotic cells, and systems for detecting membrane protein interactions [29–33].

The setup initially designed to explore pairs of proteins has been adapted to screen entire genomes in two different approaches [34–36]:

- Matrix approach. A matrix of prey clones and mated bait strains is created using distinguishable well plates. Those wells showing interacting chimeric proteins are selected based on the expression of the reporter gene.
- Library approach. The library may consist of random cDNA fragments or open reading frames representing the proteins expressed in a particular organism or tissue. Positive interactors are usually selected based on the ability of the engineered strain to grow in specific substrates.

Independently of the high throughput approach followed, it is noteworthy that Y2H is intended to detect binary interactions, even if the technique is applied genome-widely.

1.3.2 Affinity purification/mass spectrometric identification

Tandem Affinity Purification (TAP), usually combined with Mass Spectrometry (MS) techniques, is a powerful methodology to identify protein-protein interactions and, in particular, protein complexes. The TAP method involves the fusion of the target protein C-terminus with the TAP tag. This TAP Tag is a multi-domain chimeric protein containing calmodulin binding peptide (CBP) in the N-terminal, followed by the tobacco etch virus protease (TEV protease) cleavage site and Protein A, which binds tightly to IgG [37, 38]. The engineered protein is expressed in the host cell where it can form native complexes with other proteins.

The target protein and the interacting proteins are isolated using a two step purification process. First, the protein tightly binds to beads coated with IgG; after washing out the contaminants, the TEV protease cuts the cleavage site. The elute of this first step is then adsorbed in calmodulin-coated beads in the presence of calcium. After washing, the target protein complex is released using ethylene glycol tetraacetic acid (EGTA).

The elution, containing the target protein and the interacting partners, is screened by polyacrylamide gel electrophoresis, cleaved by proteases and the fragments identified by MS. The basis of MS is to produce ions that can be identified based on their mass-to-charge ratios [26, 39, 40]. MS works by ionizing the peptides to produce charged molecules in the gas phase that could be analyzed and detected using electromagnetic fields [41–43]. Different algorithms implement the identification of the resulting mass spectra [44–47]. Despite being able to find different variants of MS applied to the characterization of protein-protein interactions, purification of protein complexes still remains as the bottleneck of the process.

1.3.3 Synthetic lethality

Synthetic lethality is a very common type of *in vivo* genetic screening which tries to understand the mechanisms that allows phenotypic stability despite genetic variation, environmental changes and random events such as mutations. This methodology produce mutations or deletions in two or more genes which are viable alone but cause lethality when combined together under certain conditions [48–52]. A screening of these genetic dependencies can point to possible physical interactions between proteins, their participation in the same biochemical process or a similar function. Compared with the results obtained in aforementioned methodologies, the relationships detected by synthetic lethality not necessary requires the physical interaction between the proteins. Therefore, we refer to this type of relationships as functional interactions.

1.3.4 Protein chips

Protein microarrays or protein chips are frequently used to detect interactions, but also to determine the function of the interacting proteins - especially in large scale experiments [53–55]. The chip consists of a support surface where an array of proteins is immobilized. Probe molecules labeled with fluorescent dyes are then added to the platform. After washing the surface, any specific interactions can be noticed through the detection of fluorescent signal by a laser scanner. The main advantage of this technology is the ability to test a large number of molecules in the same experiment. Unfortunately, and contrary to what happens with DNA microarrays, the stability of the proteins is very sensitive to their environment so the performance of the platform decreases when conditions are not controlled.

1.3.5 Phage display

There are different implementations of the phage display methodology as well as different applications [56]. One of the most common protocols uses the M13 filamentous phage. The DNA encoding the protein of interest is ligated into the gene encoding one the coat proteins of the virion. Modified *Escherichia coli* strains are then transformed with the phage gene and the inserted DNA, whose products will not be released until the cells are infected with a helper phage. By immobilizing the other protein on the surface of a microtiter plate, a released phage displaying the partner protein remains, while others are removed by washing. Those remaining can be eluted and used in an iterative process to enrich the sample with binding proteins. The high throughput implementation of phage display usually implies the usage of immobilized bait and a library consisting of all coding sequences in a cell or tissue. Normally, the process is followed by computational identification of potential interacting partners and a yeast two-hybrid validation step [57].

1.4 Limitations of experimental methods

High-throughput techniques have been applied during the last years in order to systematically identify protein interactions. The resulting evidences need to be considered in the context of

the technique used to detect them, since the limitations of these technologies may influence the reliability of the candidate interaction. Some of those limitations emerged when methods as Y2H or TAP-MS were applied to determine genome-wide protein interactions. Interaction maps using these approaches have been described for different model organisms such as *Helicobacter pylori* [58], *Escherichia coli* [59], *Saccharomyces cerevisiae* [60–64], *Caenorhabditis elegans* [31, 65], *Drosophila melanogaster* [66] and *Homo sapiens* [67, 68]. However, the low reliability of this analysis is one of the main drawbacks in the large scale study of protein interactions. For example, the first two genome-wide analysis performed in yeast revealed 692 and 841 putative interactions, respectively [60, 61]. Nevertheless, the overlapping between both sets was of about 20% of the interactions [60]. More recent studies estimated a false-negative rate of 90% and a false-positive rate of 50% for these datasets [69, 70]. The reason for these particularly poor numbers could be partially explained by the unstable nature of many interactions.

In spite of the natural causes, the inherent experimental biases of the most used techniques need to be addressed. Y2H and TAP-MS, for instance, generate a lot of false positives and miss a lot of known interactions. Y2H has the advantage of being an *in vivo* technique able to accurately detect interactions without prior knowledge of the complex. However, their results are biased towards nonspecific interactions, especially if the following additional methodological limitations are not considered [71]. Firstly, an important limitation on the coverage of Y2H is related to the fact that not just any protein can be targeted. Since proteins initiating transcription by themselves produce false positives, the study of transcription factors and their interacting proteins requires alternative methods. Secondly, the structural effect of expressing sequence chimeras might be particularly awkward as fusion can change the structure of the target protein. Thirdly, the experimentalist needs to be aware that protein features such as post-translational modifications are not necessarily conserved between the organism of interest and yeast. On the other hand, even though TAP-MS can report indirectly bound proteins forming part of protein complexes, the contamination of the target is a frequent disadvantage especially if we don't have prior knowledge of the system. Therefore, the majority of experimental evidence cannot distinguish between direct interactions and those mediated by at least one intermediate protein [72]. Moreover, since the interactions are reproduced *in vitro*, the consequences of altering the protein environment and therefore the interaction are hard to predict. This effect is particularly critical on transient interactions which are particularly elusive on TAP-MS. Another limitation common to most of the experimental techniques is the fact that they hardly detect interactions involving proteins in low abundance. In consequence, even the curated datasets are heavily biased towards proteins in high abundance [69].

Although the accuracy and coverage of these techniques gets better every day, all the discrepancies found in the published datasets, together with the aforementioned technical issues, invite to be cautious when interpreting results from high-throughput studies. Accuracy can be increased by combining data sets [69, 73, 74], by repeated screening [75] and by confidence evaluation (section 1.6.6) [74]. Nevertheless, these additional steps require an additional cost added to the price of experimental procedures expensive by themselves.

1.5 Protein Interaction Databases

Several repositories of protein-protein interactions are publicly available to provide access to experimental evidences. While some databases store interactions directly submitted by experimentalists; in others, the interacting proteins are obtained by mining the literature or contain functional associations. Whether these evidences of interaction are curated manually or by automated algorithms also depends on the database (Table 1.1). The information supporting the interaction varies depending on the resources, so efforts to standardize the annotations are ongoing [76]. Indeed, different experimental techniques provide complementary information. Y2H, for instance, gives the identity of interacting proteins, while electron microscopy provides relative positional information regarding the proteins, and crystallography provides full atomic detail of interaction surfaces. In spite of the variability of the stored data, a considerable overlapping exists in the contained information [77]. In the future, further development and curation of interaction databases will be necessary.

Table 1.1: PPI Databases

Databases	Experimental	Structural	Functional	Manual Curation	Species Specific	References
DIP, LiveDIP	✓	✓				[78, 79]
BIND	✓	✓		✓		[80]
Intact	✓			✓		[81]
BioGRID	✓			✓	✓	[82]
MIPS/MPact	✓		✓	✓	✓	[83, 84]
MPIDB	✓	✓	✓	✓		[85]
HPRD	✓			✓	✓	[86]
STRING	✓		✓			[87]
ProtCom		✓				[88]
Prolinks			✓			[89]
ECID			✓		✓	[90]

1.6 Computational methods

The described limitations of the experimental techniques call for the development of new approaches to predict whether two proteins interact. The interacting partners usually share some structural, physicochemical or evolutionary constraints as a consequence of their interaction. Therefore, several computational algorithms have emerged based on different descriptors, in order to predict interactions at a large scale. While some methods are based on the structural features of the candidates, others may take advantage of the increasing genomic information available [91]. The latter, known as *genome context* methods, can provide additional evidence on candidate interactions, as well as some insight on the evolutionary events governing them (Figure 1.1). Since interactions detected by *genome context* methods relies on indirect evidences fixed by evolutionary pressures, the associations established not necessary imply physical interaction, but are involved

in similar biological functions such as the same metabolic pathway, the same protein complex or the same operon. Therefore, we usually refer to these associations as functional interactions.

1.6.1 Gene fusion events

Methods based on gene fusion events, an approach also known as "Rosetta Stone", are based on the observation that some interacting proteins have homologs in other genomes that are fused in one polypeptide chain [92, 93](Figure 1.1). For instance, the alpha and beta subunits of tryptophan synthetase are fused in fungi but exist as separate chains in bacteria [94]. Gene fusion apparently occurs to optimize co-regulation synching the relative concentration of both species. Analysis of pairs predicted by this approach revealed that for more than half of the pairs, both members were functionally related [95]. Gene fusion is particularly frequent in metabolic proteins [96].

The algorithms to detect such kind of events usually imply sequence searches and multiple sequence alignments (MSAs). More recent modifications have included statistical measures to detect *gene fusion* focusing on all homologs rather than restricting the analysis to the orthologs [97].

The main limitation of this methodology lies on the prevalence of fused proteins. By definition, this approach is restricted to shared domains in distinct proteins, a phenomenon whose true extent is still unclear [98], especially in prokaryotic organisms Moreover, the presence of promiscuous domains such as SH2 or SH3, invite to be cautious when applying this approach in an automatic way.

1.6.2 Gene neighborhood and gene cluster methods

Functionally related genes, which encode potentially interacting proteins, are frequently transcribed together as: operons in bacteria and co-regulated clusters in eukaryotes. Gene neighbor methods apply these adjacency relationships as a proxy to infer interacting proteins (Figure 1.1). Despite the gene shuffling produced as an effect of neutral evolution, gene order is usually conserved resulting in gene clusters and operons [99, 100]. By using this neighborhood information, functional interactions were predicted with higher coverage (about 37%) and precision (63–75%) than prior genomic inference methods [101]. To some extent, this approach can be applied to eukaryotes, in which interacting co-regulated genes are often found to cluster in the genome [102].

Successful examples of gene neighborhood have been reported. By comparing gene order in archeal and eukaryotic genomes, the exosome superoperon suggested a functional and possibly a physical interaction between two subunits of RNase P and the postulated archaeal exosome, a connection that had not been reported in eukaryotes [103]. This novel linkage was validated experimentally in supplementary studies [104].

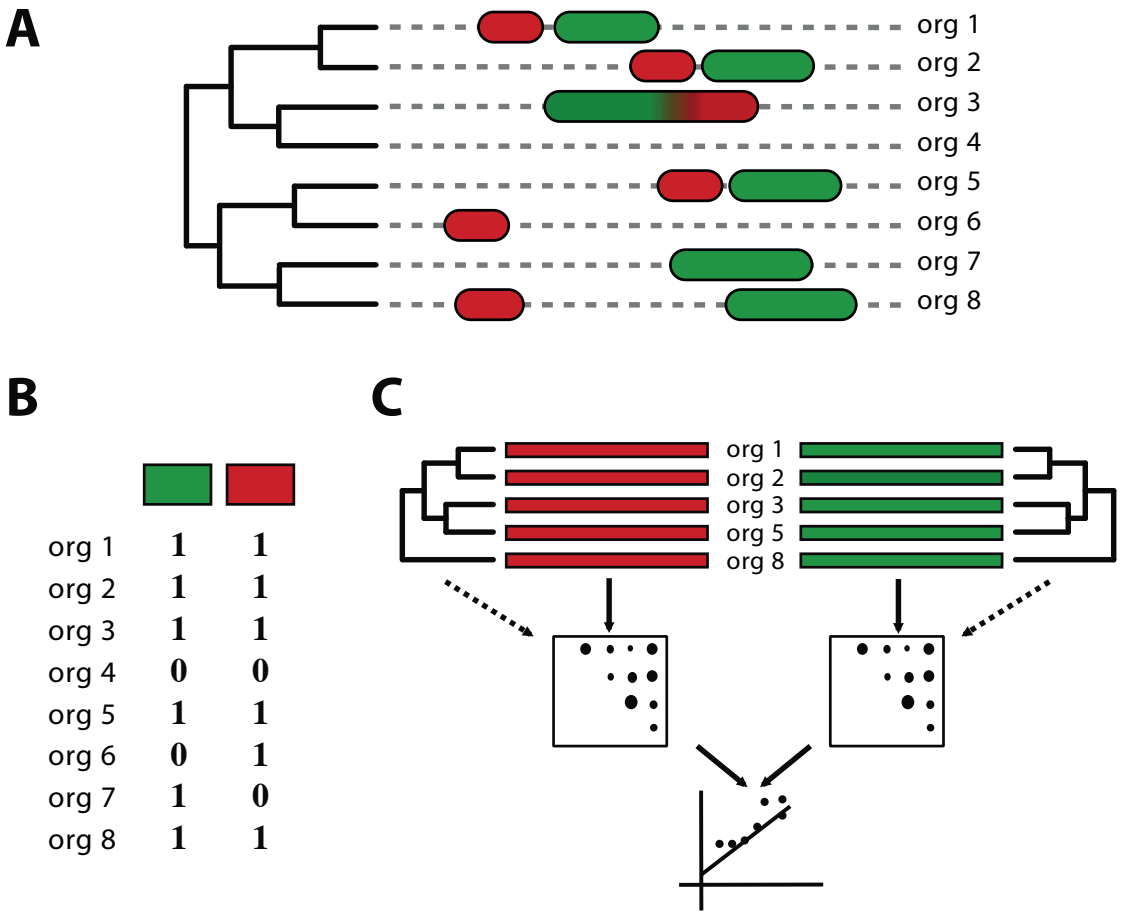


Figure 1.1: **Computational methods for predicting protein-protein interactions.** *Genomic context* methodologies predict the eventual interaction between two proteins - red and green - only using the information encoded at a genomic level. **A.** Fused genes (i.e. "org 3") may suggest interaction. In the other hand, genes in different *loci* but conserving genomic closeness (i.e. "org 1", "org 2" and "org 5") usually share some functional relationship. **B.** Phylogenetic profiles are created based on the presence/absence of certain genes in a set of genomes. Similarity of phylogenetic profiles is calculated using different metrics in order to predict interactions. **C.** Phylogenetic trees, calculated using the distances between homologous sequences, are compared in order to detect co-evolution processes that may indicate interaction.

1.6.3 Similarity of Phylogenetic Profiles

Methods based on phylogenetic profiles (PP) are founded on the observation that functionally associated and potentially interacting proteins evolve in a codependent manner, that is, they tend to be jointly inherited or eliminated. One hypothesis explains this presence/absence pattern as "reductive evolution". If the cooperative interaction of both proteins determines the function of the system and one of the proteins is lost, the evolutionary pressure to maintain the other protein disappears. Likewise, if a protein is recruited, the interacting partner needs to be "acquired" in order to form the functional complex. The underlying idea is that many complexes or pathways require all their members to be present in order to fulfill their biological function. These events are compatible with the idea of the "selfish operon" in which a cohort of genes suffer horizontal gene transfer together [105].

A phylogenetic profile is constructed for each protein as a vector representing its presence/absence in a set of genomes. Originally, these vectors were binary, where "1" denotes presence and "0" absence in a qualitative way [93, 106, 107](Figure 1.1). Nevertheless, the binary representation is a simplistic way to encode the evolutionary events of a protein in a profile. Therefore, the similarity of the sequences to the sequence of a reference organism was incorporated at each position, in order to enrich the vectors with quantitative data [108]. Phylogenetic profiles constructed at domain level [109–111] or using phenotypic traits [112, 113] have also been used in both qualitative and quantitative ways. Once the profiles are defined, the similarity is calculated by different measurements including euclidean distance [93], mutual information [108] or Hammings distance [114].

Several studies demonstrate that proteins with similar phylogenetic profiles are functionally related. For instance, homologs of *E. coli*'s ribosomal protein RL7 were found in most eubacterial genomes and in yeast but not in archeal genomes. The same pattern was obtained in many ribosomal proteins functionally related to RL7 [107]. Other examples of correlated profiles include flagellar proteins and proteins involved in amino-acid metabolism [107]. The application of phylogenetic profiles can also be extended by considering the functional linkage when the profiles are anti-correlated [115]. The assumption that relationships between functionally related phylogenetic profiles can be explained using logic operators beyond mere co-presence ("AND operator") was more recently exploited using triplets of profiles displaying higher order relationships such as complementation [89].

The performance of phylogenetic profiles is strongly affected by the set of organisms selected to build the profiles. As more and more completely sequenced genomes become available, more evident is the fact that the best predictor not necessary requires all the available data [116]. Indeed, depending on the type of functional relationship we are trying to detect, the optimal set of organisms may change. Using organisms belonging to the three superkingdoms has been proposed as adequate for detecting conserved interactions (e.g. translation apparatus), whereas species in the same superkingdom are more accurate for more variable pathways such as carbohydrate metabolism [117]. Comprehensive studies recently analyzed the effect of reference taxa using a set of 565 bacteria, suggesting that sub-samples of organisms can achieve better performances than

the whole set of available genomes [118]. Since the manual selection of the reference set could be arbitrary, machine-learning algorithms based on known protein interactions have been developed, reaching higher accuracies than their unsupervised counterparts [119].

Using the mentioned methodology, the presence/absence of every gene is equally weighted, independently of the number of causing evolutionary events. The identification of gain/loss events by combining phylogenetic trees with phylogenetic profiles allows the estimation of profile similarity likelihoods. Different implementations of this approach include the use of Markov models [120, 121], kernel trees [122] or explicit comparisons [123]. All previous approaches have not considered the fact that gene clusters may strongly coevolve in some parts of the evolutionary tree while exhibiting a very weak signal in other periods. This asymmetry on the co-evolutionary signal known as *local co-evolutionary* problem has been subject to different studies but remains a computationally challenging task [124, 125].

Phylogenetic profiles have proven to be a powerful and intuitive methodology in the last decade, though some disadvantages need to be addressed. Besides the previously mentioned limitations, one of the most important problems lies in the construction of accurate profiles. Since the profiles are usually constructed based on orthology relationships, this methodology can only be applied to complete and well-annotated genomes to be sure of the absence of a given gene. Even in those cases, the orthology assignment is not trivial, being especially critical for eukaryotes, in which the presence of pseudogenes or inactivated genes makes the orthology assignment awkward. Moreover, essential proteins or those specific for a given genome behave as hard candidates since they are encoded as flat profiles.

At the practical level, predicted interactions derived from automatically generated phylogenetic profiles are available in several resources such as STRING [87], Prolinks [89] or ECID [90].

1.6.4 Similarity of Phylogenetic Trees

As discussed previously for the phylogenetic profiles, interacting proteins very often coevolve, so that changes in one protein are associated with correlated changes in the partner. However, since phylogenetic profiles encode for dramatic events as a consequence of gene gain/loss in different species, they ignore coordinated changes that may be occurring at a sequence level. Such correlated changes were qualitatively reported to occur between some families of ligands and their receptors, resulting in similar phylogenetic trees [126–128]. The similarity between the phylogenetic trees of these protein families, despite not being quantified at that time, was interpreted as co-dependencies. However, when the genomic revolution arrived, methods to measure this similarity at a large scale became mandatory.

The first method to estimate tree similarities was based on calculating the correlation coefficient between distance matrices, as proxies of the phylogenetic trees [129]. The algorithm was soon scaled up to a genomic level under the name *mirrortree*, by using the correlation coefficient as an indicator of protein-protein interactions [130, 131](Figure 1.1). On this initial implementation, for the two protein families for which co-evolution is to be evaluated, multiple sequence alignments are generated using orthologs obtained from a set of reference genomes. Tree similarity is estimated

by calculating the correlation coefficient between the inter-ortholog distances in these alignments. Unambiguous correspondence between the sequences of the two alignments is required so as the elements of the two distance matrices can be compared. When a pair of protein families shows a high correlation coefficient, the proteins are predicted as interacting. More recent modifications suggest that when cophenetic distances extracted from the branch lengths of the phylogenetic trees themselves are used for correlation calculation, the performance of the prediction is slightly improved. [132].

Many successful applications of the *mirrortree* methodology have come out over the last years [133–139]. Correlated phylogenetic trees often appear in systems where the proteins have to change and, at the same time, maintain interactions [140]. Some studied examples are listed below.

- NADH-ubiquinone reductase. The tree similarity between the phylogenetic trees of some members of this complex is particularly high. In *Escherichia coli* for instance, the NuoE and NuoF subunits show a clear tree similarity (0.86 in a 0–1 scale) and their interaction is supported by the 3D structure of an homolog complex [141, 142].
- Peroxiredoxins. Recent reports suggest that this family of proteins experience oxidation-reduction cycles which are markers of circadian rhythms in all domains of life. The correlated evolution between the members of this family and a previously characterized clock mechanism in cyanobacterias has helped identify this novel functional role [143].

To make this and other analysis possible, a web server to predict protein-protein interactions by the *mirrortree* method was published by other group [144]. The tool automated most of the bioinformatic workflow in order to allow non-expert users to perform co-evolution analysis. The user provided a pair or a list of protein sequences and pairwise comparisons are calculated. Although the user could set different thresholds, no other interactive features existed for tree exploration. Moreover, the server is no longer available.

1.6.4.1 Co-evolution and Co-Adaptation

As mentioned above, phenomena such as similarity of phylogenetic trees, and similarity of phylogenetic profiles are indicative of co-evolution. However, the meaning and implications of this concept at a molecular level as well as the ultimate cause for this observed co-evolution have not been adequately addressed so far. At the species level, co-evolution is a well-documented phenomenon involved in the organization of biological communities and a relevant component of the current evolutionary theory. All these ecological concepts and methodologies applied at a species level can easily be extrapolated to the study of interacting genes and proteins.

Although the initial ideas of co-evolution at species level can be traced back to Darwin's work (1862) on orchids and pollinators [145], the term co-evolution is attributed to Ehrlich and Raven (1964), during their study of the reciprocal evolutionary relationships of butterflies and their food plants [146]. The most widely accepted definition was stated by Thompson as the "reciprocal

evolutionary change in interacting species" [147]. The importance of this definition lies in the reciprocity. The changes in one population imply changes in the selection pressure of a second population and vice versa. Several ecological examples of co-evolution at the species level have been described over the years, including inter-specific competition for resources, relationships between parasites and their hosts, predators and preys or symbioses [148]. In some of these cases, morphological traits emerged in a correlated fashion as a consequence of the co-adaptation between the interacting species [147]. Hence, phylogenetic trees of co-evolving species usually share some topological similarity, for instance in the case of the parasites and their hosts [149, 150]. However, not every similarity should be understood as a direct influence between the interacting species. Indeed, as species evolve in response to a complex environment formed by multiple ecological factors and interactions, a background force related with the constant improvement of the fitness remains common for related species. As a consequence, the term "diffuse co-evolution" or "guild-co-evolution" was coined in order to refer to those cases where the mutual influence of the co-evolving species cannot be demonstrated [147, 151]. The constant improvement in the fitness of species in an ever-changing environment was formulated as the "Red Queen Hypothesis" [152–154]. On the other hand, the term co-adaptation refers to the coordinated changes responsible for the specific mutual adaptation of species [155–158].

All these concepts can easily be transferred to a molecular level: a change in one locus alters the selective pressure of another locus, and this change is reciprocal [159]. These co-dependencies have been observed within different residues or regions of the same molecule, or between different molecules such as interacting proteins. Similarly to co-evolution at species level, one important question that arises is to what extent the observed co-evolution is due to compensatory changes in the interacting loci (co-adaptation) or to general factors that affect both locus in a similar degree. In the evolution of proteins in the same complex for instance, we may expect some similarities as a consequence of the mutual adjustments necessary to maintain the interaction. In the same way, certain correlated changes may occur to preserve co-expression, foldability and all other constraints that are imposed to preserve functional properties of the interacting partners. It is important to note that some of these factors are common even for proteins that do not directly interact but share some general biological functional. Therefore, techniques such as *mirrortree*, which are based on correlated distances rather than on a direct measurement of any of the aforementioned factors, will hardly distinguish which one of them is the dominating contributor to the signal.

Co-evolution and co-adaptation are ongoing processes shaping the phylogenetic trees of interacting proteins. Given that co-evolution at molecular level may appear as a consequence of many functional dependencies, we expect that the predictions obtained measuring protein co-evolution might also evidence interactions at a functional level. However, similarity of phylogenetic trees is also affected by the less-frequent co-adaptative events between interacting proteins. For that reason, the ability to disentangle co-adaptation from unspecific co-evolution will determine the success of computational methods to predict physical or functional interactions.

1.6.4.2 Correcting background similarity

One of the main problems of the original *mirrortree* algorithm was the large number of false positives it produced. Many pairs of proteins whose interaction was not described displayed high correlation coefficients, considerably reducing the applicability of the methodology. One of the possible reasons for such similarity between unrelated proteins can be associated with the background speciation events underlying both trees. As both partners are influenced by the speciation process, we can assume that, independently of their functional role, the trees will share a certain basal similarity with the canonical tree of life. Therefore, several methods emerged to exclude the information about the phylogenetic relationships of the reference genomes, in order to compare the residual information.

Different statistical techniques, as well as different representations of background similarity, have been developed for "phylogenetic subtraction". The first attempts to remove the speciation signal subtracted 16s rRNA phylogenetic distances directly from the matrices of the interacting candidates [132, 160]. This corrected methodology, known as *tol-mirrortree*, found a real interactor among the 6.4% top scores for half of the proteins, compared with the 16.5% obtained by the original *mirrortree*. More successful examples of ligand-receptor interactions exist, this time applying a background speciation correction [161]. However, the direct subtraction of distances from the matrices ignores the underlying phylogenetic dependencies between the ortholog sequences. As a consequence, some sophisticated methodologies tried to calculate the corrected similarities by aligning high-dimensional embeddings of the trees [162].

Instead of using canonical trees to address the problem of unspecific tree similarities, some studies suggested inferring that background signal from the tendencies observed in large collections of protein families. These methods are based on the idea that unspecific similarities within a pair of phylogenetic trees can be deduced by comparing them with many others. One of the first attempts to take advantage of this contextual information introduced a partial correlation coefficient as a measure of similarity. This metric calculates the correlation between a pair of phylogenetic vectors, excluding the information of a third vector which contains the background information. By using the variability of the phylogenetic data as third vector, the false positive rate was drastically reduced [163]. An important improvement of this approach has been the use of the genome-wide co-evolutionary network obtained from the pairwise comparison of the proteins to remove background similarity. This approach, called *ContextMirror*, uses the comprehensive calculation of *mirrortree* similarities for all pairs of proteins in a given organism to calculate the partial correlation of a pair of proteins using a third protein as background correction. Since this third protein can be selected from a big set of proteins, the results are ordered based on the level of similarity with the pair of vectors under study. The level obtained determines the specificity of the signal acting as background, so it provides more insight on the nature of the co-evolution. *ContextMirror* displays performances comparable to some experimental techniques. In fact, context based methodologies have proven particularly accurate in reconstructing large machineries such as the bacterial flagellum or the previously reported NADH-quinone oxidoreductase [164].

1.6.4.3 Patterns of taxonomically local Co-evolution

The *mirrortree*-related methodologies usually assume that the co-evolutionary signal shared by a pair of proteins spreads over the entire tree of life. However, as we review in the case of phylogenetic profiles (section 1.6.3)), interacting proteins may strongly coevolve in some parts of the evolutionary tree while exhibiting a very weak signal in the rest of the species. Local similarities may be related with recent interactions, whereas global similarities in the phylogenetic tree may evidence a relationship occurring since ancestral species. Dealing with this non-homogenous nature of the co-evolutionary signal is not trivial as it raises certain combinatorial problems.

MatrixMatchMaker handles this problem by finding the largest common sub matrix compatible with certain evolutionary distance matrices. This bottom-up approach only looks for those sequences most strongly implicated in the co-evolutionary signal. The matrices can therefore include erroneously assigned sequences or paralogs, since these will most likely be excluded from the final similarity. By playing with a tolerance threshold, the extent of the co-evolutionary signal can be weighted. Within the framework of this approach, proteins that are known to interact in humans showed a more strong signal than proteins that simply belong to the same biochemical pathway [165]. Indeed, the human co-evolutionary network reveals two topological partitions, one generally representing ancient eukaryotic functions, and the other, modern functions acquired during animal evolution. The latter is enriched in proteins involved in pathology-related functions such as multicellularity, cell division or cell communication [166]. Further implementations have provided considerable speedups on computational time [167].

In the *mirrortree* derived approaches commented so far, the right mapping between the sequences of both candidates needs to be known in order to calculate their tree similarity. However, when duplicated genes are present, deciphering the corresponding ortholog in the other tree can be a challenging task. Under the hypothesis that the correct mapping maximizes the similarity, several methods emerged to detect interacting partners within large families of duplicated genes [117, 168–172].

The problem of selecting an optimal reference set of organisms and their influence on interaction prediction remains unclear, though. Considering the increasing number of completely-sequenced genomes, the search for local evolutionary signals would not be enough as the problem grows exponentially with the number of sequenced organisms. In the future, the selection of representative subsets of organisms might be valuable to predict type-specific interactions.

1.6.4.4 Structural features

Although there is an area of research focused on using three dimensional data to identify protein interactions [173], the genome-wide analysis would require structural data for every single protein in the organism. Alternatively, the addition of available structural information to the predictions based on similarity of phylogenetic trees can outperform the current methods as well as reveal more insight on the causes of this similarity. For that reason, different studies have focused on whether the co-evolutionary signal is homogeneously distributed along the protein sequences or enriched in regions with structural relevance for the interaction.

Instead of using whole-protein sequences, phylogenetic trees derived from the protein domains involved in the interaction display greater similarity than non-interacting domains within the same proteins. For example, by comparing the domains of the alpha and beta subunits of the mitochondrial F1-ATPase, seven out of the nine domain pairs that are known to interact present a higher correlation than the two non-interacting pairs [117]. Despite the structural complexity of the ATPase, it is remarkable how the co-evolutionary signal of the interacting partners remains informative after splitting into their constituent domains. As in the case of whole protein sequences (section 1.6.4.2), the domain interaction prediction is significantly improved by removing the background similarity of phylogenetic trees. Indeed, the predictor shows more accuracy when the background removal is applied to the trees based on the most conserved residues, suggesting that both signals are more easily disentangled on those regions [174].

The similarity of phylogenetic trees at the domain level suggests the presence of local regions which not necessarily share the evolutionary constraints of the whole protein. Although different studies have tried to explain whether different structural features contribute to the observed tree similarity, the relative importance of the considered regions in the co-evolutionary signal remains unclear. In that sense, the effect of taking into account the protein interfaces for the co-evolution-based interaction prediction has been subject to study with contradictory results. Some evidence suggests that residues in the interfaces of stable interactions evolve at a relatively slow rate, allowing them to coevolve with their interacting partners. In contrast, the residues involved in transient interactions present higher plasticity, leaving little or no co-evolutionary signal in the interface [175]. When the residues in the interfaces are removed, the remaining sequence still contains the co-evolutionary signal necessary to predict the interaction [176]. However, whereas some authors understand these results as clear evidence that the co-evolutionary signal is uniformly spread over the sequence, others highlight the presence of strong local signals of co-evolution. The former have demonstrated that no additional improvement could be achieved on protein interaction prediction by limiting the study to either the protein surface or the interaction interface [177]. The latter have defended that binding sites together with their spatially surrounding residues provides a stronger co-evolutionary signal than the same number of randomly selected residues outside the binding neighborhood [176].

The actual implications of this debate transcend the analysis of protein interfaces and highlight the evolutionary forces governing the interacting proteins. As commented previously (section 1.6.4.1), the possible explanations of the observed tree similarities between interacting proteins range from specific co-adaptation between the partners, to general global similarities between their evolutionary rates. The co-adaptative hypothesis assumes that inter-protein compensatory changes occur in order to stabilize mutations at the interface of one interacting partner. Alternatively, some similarity between phylogenetic trees of interacting proteins can also be expected as they share similar evolutionary constraints and, therefore, similar evolutionary rates. Indeed, direct [177–180] and indirect [181–186] relationships have been established between similar evolutionary rates and protein interactions. The relative prevalence of both phenomena is subject to certain debate. Even if there is evidence of co-adaptative changes, it is difficult to think that co-adaptation is the only process involved in the observed co-evolution. Since many compensatory

changes would be necessary in order to affect sequence distances, and hence the phylogenetic trees, it has been proposed that a large proportion of the observed similarity is due to similarities in evolutionary rates. On the contrary, the fact that tree-tree similarity decreases rapidly with the distance in the protein interaction network [187], as well as other evidences [165] suggest that directly interacting proteins co-evolve stronger, supporting the co-adaptative hypothesis. Further progress on incorporating structural data on co-evolution analysis can provide a better understanding of both processes, maybe focusing on regions with large proportions of highly co-adapting residues.

Structural information is necessary and critical in order to fully understand interactions at the molecular level. Because the set of available structures persists as the limiting resource, the mentioned analyses are all based on relatively small and eventually biased groups of proteins. Moreover, the range of applicability of the suggested modifications drops drastically compared with the sequence based methodologies. Comprehensive structural-feature prediction using sequence information might bypass the problems derived of the data scarcity. Nevertheless, little is known about the effect of using predicted structural features in protein interaction prediction.

1.6.5 Supervised Methods

Other methodologies include predictors based on training with the available information on proteins and interactions. The strength of these data-mining approaches is the integration of multiple information such as gene expression, experimentally determined protein interactions, protein localizations, protein phylogenetic profiles or similarity of phylogenetic trees [188–193]. Although works using different algorithms and different organisms have been reported over the past few years, these methods suffer from the inherent problems derived from their primary data. Moreover, these methods require experimental data to be trained, contrary to the methods described so far which can be considered *ab initio* in this regard.

1.6.6 Assessing predictive accuracy

Different approaches exist for evaluating the results of a method to identify protein interactions, either experimentally or computationally. The simplest approach for evaluating a set of putative interactions is to measure its overlap with a gold standard reference set. Previous studies or protein interaction databases (section 1.5) are frequently used as gold standard [69, 72]. Another approach is to use the biological properties of candidate proteins such as protein function, localization or expression to assess how likely the interaction is to occur [62, 194–196]. Indeed, some methods combined all these features in order to give a more confident predictor of the interaction reliability [197–199]. In some other studies, interologs have been used for validation [194, 200]. Usually interologs are defined as homologous interactions: if proteins A and B interact in species S, then proteins A' and B' in species S', which are orthologs of A and B, are predicted to interact. The usage of interologs for validation have been a matter of certain debate as physical interactions

seem to be more conserved within species than across species [201]. Since all the protein interaction networks share the same topological features (section 1.1.1), methods based only on the characteristics of the resulting network have also been proposed for validation [202–204].

The computational methods, contrary to their experimental counterparts, provide a score associated with the interaction confidence. Thus, the ranked lists of candidate interactions are usually validated using both, a positive and a negative gold standard set. One of the most common approaches is the Receiver Operating Characteristic (ROC) analysis [205–207]. The true positive sets necessary to run this analysis can easily be obtained from manually curated databases, though some cautions are required in those organisms with low experimental coverage [200]. However, a negative reference set can be hard to define as the experimental data reporting non-interacting proteins is comparatively small [208]. Recently, a new database has begun to archive this data [209].

1.7 Open problems

Considering the aforementioned limitations associated with the experimental identification of interacting partners (section 1.4), the computational techniques introduced over the past years are a powerful alternative to suggest protein interactions. In particular, those methods based on co-evolution at the sequence level performed as the most promising ones, achieving genome-wide coverage and accuracies close to their experimental counterparts. However, there are still a large number of scientific and technical difficulties as many of the factors influencing the co-evolution of interacting proteins remain unknown or hard to investigate.

In that sense, different factors make the multi-step workflow necessary to perform a simple tree comparison for a pair of proteins difficult to apply by non-expert users. Firstly, the user needs to generate the phylogenetic trees of the pair of families of interest using a pipeline formed by a number of sequence analysis methodologies. Secondly, the distances of the resulting phylogenetic trees need to be compared using approaches that usually are distributed as command-line tools or need to be implemented by the user himself. Finally, the user requires a high level of expertise to put in context the results obtained, considering the complex taxonomic distribution of the protein families under study. Thus, in order to analyze a single pair of proteins, the user require a sophisticated bioinformatic setup and a deep knowledge on the tools and concepts usually applied on sequence analysis. A framework which allows to automatically generate phylogenetic trees and interactively analyze the tree similarities would help to extend the co-evolution based prediction of protein interactions to the part of the scientific community not familiarized with these techniques.

The analysis of which set of organisms or taxonomic levels optimize the co-evolution-based protein interaction prediction is also an interesting question at a systematic level (section 1.6.4.3). It is accepted that, even in the cases of tight co-evolution, the codependence between the interacting partners may not be constant over the whole tree of life. Although recent studies have described the implications of the reference set of organisms in the comparison of phylogenetic profiles, little is known about their influence when comparing phylogenetic trees. Different factors such as the

age or type of interaction might influence the interaction prediction, so a deeper understanding of the phenomenon might improve the current predictions.

An additional difficulty is related to the proper disentanglement of the clearly observable phenomenon of co-evolution and the more elusive co-adaptation (section 1.6.4.1) [141, 210]. A number of approaches have been applied for removing the unspecific tree similarity (section 1.6.4.2) by excluding the speciation events or by using the context defined by all the co-evolutionary signals in the organism. Although these approaches have helped the understanding of the functional relationships between pairs of proteins, there is no evidence that real co-adaptation is being detected. Even in protein interfaces (section 1.6.4.4) where co-adaptation is expected to occur at a higher rate, contradictory results increase the uncertainty on the extent to which the co-adaptative process is taking place. The contradictory results could be in part due to the fact that these studies combining co-evolution with structural features have been restricted to the relatively small set of available structures, limiting the true extent of these observations. Consequently, it remains to be explored the results of extending these studies with predicted structural features (in principle available for any protein).

Considering the limitations described, the correlation coefficient between inter-protein distances used as score can hardly describe the co-evolution between a pair of proteins as it is highly influenced by the set of organisms used to generate the phylogenetic trees. The more genomes are experimentally sequenced, the more the presence of closely-related organisms increases sequence redundancy. This redundancy might artificially increase the correlation coefficient, negatively affecting the method's performance. Besides, the pernicious influence of the redundant information is asymmetrically distributed along the tree of life. This phenomenon implies that, depending on how spread in the taxonomy the pairs of proteins under study are, sequences belonging to particularly redundant organisms may be considered for correlation or not, significantly changing the correlation coefficient. Thus, it is necessary to establish a estimator of the statistical confidence of the co-evolutionary process by considering the set of reference organisms and the number of organisms where the proteins are present. So far, the significance of a given correlation coefficient has been evaluated based on the number of leaves in common between the pair of trees. Using the number of organisms in common and the correlation coefficient, the statistical confidence of that result was assigned based on tables of *P*-values derived from a general null model not specifically designed for the particular case of comparing evolutionary distances. For example, that model assumes independence between the values which does not hold for sequence-based distance matrixes. We know that the phylogenetic trees cannot randomly change to acquire any possible distance between their leaves. The protein sequences can only change in a limited universe of possibilities constrained by their folding or function, among many other factors. Moreover, the distance between two leaves in a tree can not freely change without affecting others. Therefore, a revisited version of the current *P*-values, considering the space of possibilities of the sequences under study, would help the understanding of the co-evolutionary signal.

Motivation and Goals

Considering the current state of the art in the computational prediction of protein-protein interactions, the motivation of this thesis is to contribute to the improvement of the so-called co-evolution-based methods. During the last decade, this family of techniques has demonstrated its ability to perform genome-wide protein interaction predictions even reaching accuracies similar to their experimental counterparts. However, many other limitations have appeared over the years, derived from technical issues or inherited from the lack of understanding of the underlying evolutionary process. Over the coming years, while the number of totally sequenced genomes increases, these problems may have a dramatic impact on the global prediction. Therefore, we propose a study of the area from different perspectives in order to broaden our understanding of the occurring mechanisms and provide solutions to some of the current hot topics.

2.1 Objectives

This thesis can be divided into the following major objectives:

1. *Developing the Mirrortree web server to study the co-evolution of protein families*. This tool will provide a user-friendly interface allowing non-expert users to perform tree-tree comparisons in order to look for protein-protein interactions or other functional relationships. The tool must combine the automatic generation of phylogenetic trees with a proper visualizer where further exploration of the similarities can be carried out. Additional options for the more advanced users need to be taken into account.

2. *Incorporating information on predicted solvent accessibility*. The effect of including predicted structural features such as predicted solvent accessibility has not yet been analyzed yet. Here, we will asses for the first time the effect of including this information in the results and the range of applicability of three co-evolution-based methods. A proper comparison taking into account the different methods and the different kinds of interactions may help to get more insight on the co-evolutionary processes occurring on protein interactions.

3. *Studying the effect of the reference set of organisms used for building the trees on the performance of the methods*. The organisms used as reference on the different *mirrortree*-like methodologies may have a direct implication on the prediction of protein interactions. In order to asseses which are the optimal organisms for each method and type of interaction, a number of reference sets based on different taxonomical properties need to be properly benchmarked, so as to end up in a number of recipes on which organisms are adequate for each situation formulated in taxonomic terms.

4. *Predicting protein interactions on the basis of an evolutionary consistent model*. Considering the impact of the growing number of available genomes on the performance of the co-evolution-based prediction of protein-protein interaction, methods dealing with large numbers of redundant genomes need to be developed. Hence, we will work on a new methodology based on the calculation of statistically consistent *P*-values associated to the correlation scores. To calculate these *P*-values a null evolutionary model will be created based on the permutations of the phylogenetic trees under study.

Methodologies

3.1 General methods and resources

3.1.1 Co-evolution-based methods for protein interaction prediction

This family of methods compares phylogenetic trees in order to predict protein-protein interactions. The tree generation procedure varies depending on the setup of the experiment and will be explained in the corresponding sections below. Instead of using the trees themselves for comparison, distance matrices are calculated by adding the branch lengths separating each pair of proteins in the tree. The distance matrices are used as input in the following methodologies.

3.1.1.1 Mirrortree

The original *mirrortree* (MT) approach (Figure 1.1) calculates the similarity between two distance matrices by calculating the linear correlation coefficient between the corresponding values according to the standard equation [130]:

$$r_{RS} = \frac{\sum_{i=1}^{n}(R_i - \bar{R}) \cdot (S_i - \bar{S})}{\sqrt{\sum_{i=1}^{n}(R_i - \bar{R})^2}\sqrt{\sum_{i=1}^{n}(S_i - \bar{S})^2}} \tag{3.1}$$

Here n is the number of elements in the matrices, that is, $n = \frac{(M^2-M)}{2}$, M being the number of organisms in common between both trees, R_i are the elements of the first distance matrix (the distance among all the proteins in the first phylogenetic tree), S_i is the corresponding value for the second matrix and $\bar{R}$ and $\bar{S}$ are the averages of R_i and S_i respectively.

In most of the implementations, correlations between a pair of phylogenetic trees are only considered if they share at least 15 species. Moreover, the statistical confidence of the obtained result is usually calculated based on the correlation coefficient and the number of organisms in common. The more organisms in common and higher correlation, the more significant the correlation is considered. The *P*-values are obtained from pre-calculated tables of significance for linear correlation calculation.

3.1.1.2 Profile-correlation

The Profile Correlation (PC) method [164] takes the mirrortree correlation coefficients for all pairs of proteins in a given organism as input. This co-evolutionary network can be represented as a squared matrix the size of the proteome with some missing values for those pairs without a significant number of organisms or adequate *P*-value. A row, or a column, in this matrix, known as "co-evolutionary profile", represents the co-evolutionary behavior of a protein with the rest of the proteome. In the PC method, the similarity between a pair of proteins is reassessed as the correlation between their corresponding co-evolutionary profiles r'_{RS}:

$$r'_{RS} = \frac{\sum_{i=1}^{N}(r_{Ri} - \bar{r}_R) \cdot (r_{Si} - \bar{r}_S)}{\sqrt{\sum_{i=1}^{N}(r_{Ri} - \bar{r}_R)^2}\sqrt{\sum_{i=1}^{N}(r_{Si} - \bar{r}_S)^2}} \tag{3.2}$$

Here, N is the number of proteins in the genome for which the correlation values with both R and S could be calculated. r_{Ri} and r_{Si} represents the *mirrortree* correlation coefficients (Equation 3.1) between the proteins R (and S) and the rest of the proteins in the genome. $\bar{r}_R$ and $\bar{r}_S$ are the averages of r_{Ri} and r_{Si} respectively. For PC, the same significance thresholds used for correlation in the original *mirrortree* are applied. The general idea behind PC is that we can use of the information contained in the whole "coevolutionary network" of an organism (the network containing all of the pairwise tree similarities) to gain information on the "coherence" or robustness of a given coevolutionary signal.

3.1.1.3 Context-Mirror

The ContextMirror (CM) score for a pair of co-evolutionary profiles (proteins) is calculated as the partial correlation between both profiles, taking into account a third one so as to remove the unspecific similarity explained by it [164]. The partial correlation is calculated as follows:

$$\rho'_{RS.Z} = \frac{r'_{RS} - r'_{RZ} \cdot r'_{SZ}}{\sqrt{(1 - r'^2_{RZ}) \cdot (1 - r'^2_{SZ})}} \tag{3.3}$$

Where $\rho'_{RS.Z}$ is the partial correlation between the co-evolutionary profiles of proteins R and S, holding the co-evolutionary profile of protein Z constant, and r'_{RS}, r'_{RZ} and r'_{SZ} are the PC scores (Equation 3.2) between the co-evolutionary profiles of R and S, R and Z, and S and Z, respectively.

By this approach, it is possible to disentangle the specific signal between two co-evolutionary profiles from the unspecific similarity represented by the third profile. Since the third protein could be any other protein in the organism, the results are arranged by different levels of specificity, starting with the co-evolutionary profile most similar to the pair of profiles under study.

3.1.2 Datasets of protein interaction and functional relationship

In order to evaluate the performance of the tree-similarity based methods for predicting protein interactions, both positive and negative datasets of interactions are required for the organism of interest. Here, we used three different gold standard datasets containing different types of protein interactions for the model organism *E. coli*:

- Binary physical direct interactions obtained from MPIDB [85]. This database contains binary physical interactions manually curated from the literature or imported from other databases. This version of the database contains 2,103 binary interactions between 1,538 different *E. coli* proteins. This dataset will be referred to as "Binary physical".
- Physical interactions inferred by co-presence in the same macromolecular complexes. This physical interactions may be direct or involving some intermediate proteins. The protein complexes are experimentally determined and extracted from the EcoCyc databases [211]. The set includes 1,354 pairs between 591 proteins. This dataset will be referred to as "Complexes".
- Functional interactions inferred as co-presence in the same metabolic pathways. Interaction evidence are also retrieved from the EcoCyc [211]. This dataset comprises 4,491 pairs between 719 proteins. This dataset will be referred to as "Pathways".

A summary of the characteristics of these datasets can be found in Table 3.1. Whereas the last two resources were previously used on interaction validation by Juan et al. [164], the "Binary physical" dataset provides some additional information, as it only contains curated direct physical interactions. For each positive dataset, the corresponding negative set was generated using all the pairwise combinations between the proteins involved in some positive pair, excluding those already reported as interacting.

Table 3.1: Protein interaction datasets

Dataset	Pairs	Proteins	Database	Reference
Binary physical	2,103	1,538	MPIDB	[85]
Complexes	1,354	591	EcoCyc	[211]
Pathways	4,491	719	EcoCyc	[211]

3.1.3 Performance evaluation

The aforementioned methodologies (section 3.1.1) produce large lists of protein pairs sorted by the score of the corresponding method. Additionally, taking into account the reference sets of interactions (section 3.1.2), these protein pairs can be labeled as positives or negatives. A particular list will represent a better predictive power if the positives tend to cluster at the top of the ranking and worse if the positives and the negatives are randomly ranked. The question about whether to evaluate these lists is not trivial. Here we describe different approaches used for evaluation.

The "receiver operating characteristic" analysis (ROC [205]) illustrates the performance of a binary classifier as its discrimination threshold is varied. The ROC analysis generates a plot of "true positives rate" (TPR) against "false positives rate" (FPR) when varying the score of the predictor. Curves above the diagonal represent methods with some discriminative power, being more discriminant when the curves get close to the top-left corner of the plot. Therefore, the Area Under the ROC Curve (AUC) is usually calculated as a quantification of the global performance of the prediction, ranging from 0.5 (random classifier) to 1.0 (perfect classifier).

In order to compare the ROC curves when the predictions contain different numbers of pairs, FPR and TPR are calculated respective to the total number of positives and negatives in the gold standard dataset. The curves defined this way, also called partial-ROC curves, provide not only an idea of the ability of the method to separate positives and negatives, but also of its range of applicability: whereas longer curves represent results with more predicted pairs, shorter curves are based on fewer observations.

Both ROC curves and partial-ROC curves are generated by cutting the sorted list of scores at different thresholds and plotting the resulting TPRs against FPRs. The difference relies entirely on the way TPR and FPR are calculated:

$$TPR = Sensitivity = \frac{TP}{P} = \frac{TP}{TP + FN} \tag{3.4}$$

$$FPR = 1 - Specificity = \frac{FP}{N} = \frac{FP}{FP + TN} \tag{3.5}$$

TP, FP, TN and FN are the true positives, false positives, true negatives and false negatives according with a given threshold. The positives and negatives, P and N respectively, vary depending on whether we calculate the regular ROC curve, in which we calculate them from the pairs under evaluation, or the partial-ROC curve, in which case the positives and negatives are calculated from the dataset used for validation. Note that both parameters can also be interpreted in the context of "sensitivity" and "specificity", as indicated by the equations.

ROC analysis measures the global ability to distinguish the positives from the negatives. In some cases, we may be only interested in predicting a few number of highly-confident interactions. Thus, an alternative analysis more focused on the top scoring pairs and on the positives needs to be performed. Another possible way of evaluating a sorted list focused on the positives is using "Precision" and "Recall". Considering a cut in the list of predictions produced by a given threshold, "Precision" (also known as "positive predictive value" - PPV) and "Recall"are defined as:

$$Precision = TP/(TP + FP) \tag{3.6}$$

$$Recall = TP/P \tag{3.7}$$

Usually, "Precision" and "Recall" are combined into a single parameter called "F-measure", defined as the harmonic average of both parameters:

$$F - measure = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{3.8}$$

Usually, the "F-measure" is represented against the scores of the predictor so as to detect the threshold with the best tradeoff between "Precission" and "Recall". The F-measure, obtained at this threshold, can be compared between different predictions obtaining an estimator of which is the best predictor in optimal circumstances.

3.2 Mirrortree web server

The Mirrortree web server implements the widely used *mirrortree* algorithm in a public and interactive resource for analyzing the tree similarity between two protein families. The workflow can be divided into two different parts. On one hand, starting from a pair of single sequences or a pair of aligned families provided by the user, a bioinformatic pipeline automatically generates a reliable pair of phylogenetic trees. On the other, a user-friendly interface provides different tools to explore the similarity of these phylogenetic trees or any pair of trees provided by the user. The server options and interfaces have been adapted to users with any level of expertise in Bioinformatics and sequence analysis.

Regarding the technical details, an Apache server hosts the application using PHP to register the incoming jobs. By default, server usage is limited to 1 job per computer each 10 minutes. When a submitted job starts, a Perl script is in charge of the preprocessing steps. When the results are ready, the MirrorTree Server User Interface (UI) shows up as an Adobe Flash application running on the client side. Additional sequence information is dynamically retrieved from Uniprot [212].

3.2.1 Automatic generation of phylogenetic trees

The process can be initiated from 2 protein sequences, 2 MSAs or 2 phylogenetic trees uploaded by the user through a web interface. Whereas the protein sequences start the workflow from the beginning, the MSAs and the phylogenetic trees are introduced in some intermediate steps of the pipeline (Figure 3.1). Any of those inputs must fulfill some requirements on information content and format, conveniently explained on the server's tutorial.

The Mirrortree server workflow is shown in Figure 3.1. When the user submits a pair of sequences, the first step is to reconstruct the protein families by finding all the protein orthologs that can be unambiguously related with the provided sequences. In order to do so, the input sequences are aligned using BLAST [213] against an updated version of the database of completely annotated genomes Integr8 [214]. The ranked aligned sequences are filtered by % identity, % coverage and *E*-value, to get a list of candidate homologs. By default, the filtering parameters are set on $\geq 30\%$, $\geq 60\%$ and $\leq 1 \times 10^{-5}$, respectively, but advanced users can modify them during

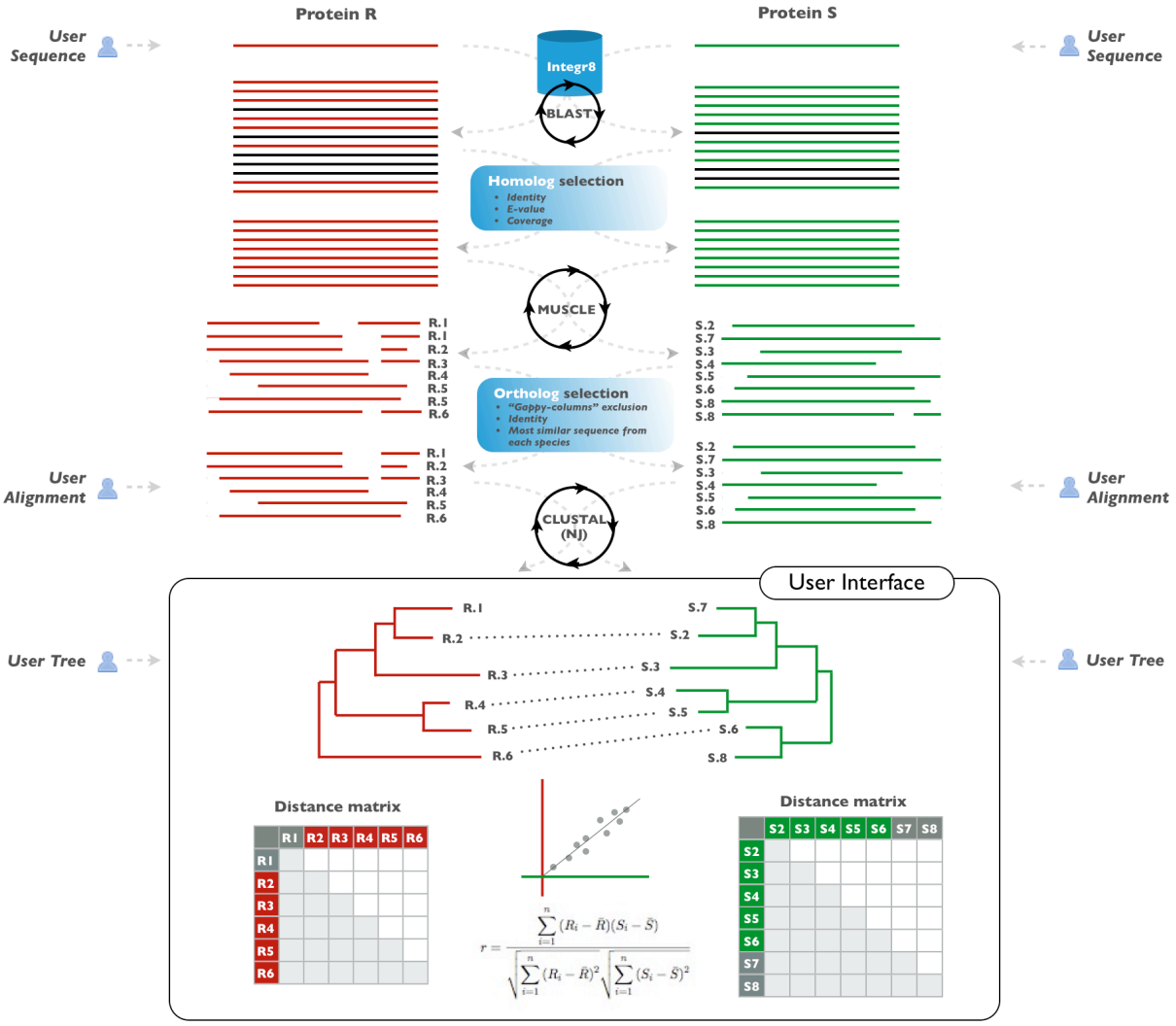


Figure 3.1: **Workflow of the MirrorTree Server.** The server automatically reconstructs the phylogenetic trees of proteins R and S based on a pair of single sequences or a pair of MSAs provided by the user. If protein sequences are submitted, they are queried against the Integr8 database, in the search for homolog sequences. After some filtering, the resulting sequences are aligned using MUSCLE to generate a pair of MSAs. After some additional filtering and ortholog selection, phylogenetic trees are generated. Once in the user interface, the correlation coefficient is calculated from the equivalent cophenetic distances in both generated trees or those provided by the user.

the job submission. The resulting lists of candidate homologs are aligned using MUSCLE [215]. To avoid server overload, this step may be delegated to a computer cluster in periods of high computing demand. From that moment on, pairwise sequence identities are calculated ignoring positions with more than 90% of gaps, in order to avoid problems with poorly aligned sequences. By using this modification, sequences with less than 30% of identity with the submitted sequence are removed from the MSAs. Finally, to gather a single putative ortholog for each organism, the sequences with the highest identity percentage with the master are chosen among the different paralogs within each organism. The two resulting MSAs have a single sequence for each organism and they may be replaced by a pair of equivalent MSAs supplied by the user (Figure 3.1). A pair of phylogenetic trees is modeled from these alignments, with the neighbor-joining algorithm implemented in ClustalW [216], using bootstrap and excluding the gaps. Both trees can also be replaced by a pair of correctly labeled trees provided by the user.

3.2.2 MirrorTree Server User Interface

When the processing of the uploaded data is finished, the user receives an email with the link to the results website. There, a number of links point to the intermediate files created during the processing, such as input files, BLAST outputs, MSAs or phylogenetic trees. Additionally, the user will find the MirrorTree Server UI. This tool post-processes the generated phylogenetic trees, estimates their similarity and plots both paired trees, allowing a more interactive insight into tree similarity.

The distance matrices necessary to perform the tree-tree comparison are calculated by the UI adding the branch lengths that separate each pair of proteins in the trees. These distances, also known as “cophenetic distances”, are calculated only for those proteins encoded in genomes present in both trees. The tree similarities, as well as the statistical confidence, are calculated based on the standard *mirrortree* algorithm (section 3.1.1.1).

3.3 Incorporating information on predicted solvent accessibility

In this section, we describe the procedure for evaluating the effect of incorporating information on predicted solvent accessibility on the *mirrortree*-based prediction of protein interactions. A detailed benchmarking, using interactions of different nature and three different *mirrortree*-related methodologies, provided more insight on the contribution of predicted accessibility to the interaction prediction. For each *E. coli* protein, predicted accessibility was calculated based on a MSAs of homolog sequences. Positions below certain thresholds of accessibility in the MSAs containing ortholog sequences are excluded. The resulting MSAs, which contained only the residues predicted as exposed according with different criteria, were used to create the phylogenetic trees as usual. The phylogenetic trees were then used for predicting interactions with three different *mirrortree*-related methodologies and the results evaluated using three different datasets representing interactions of different nature. The workflow is illustrated in Figure 3.2 and details follow.

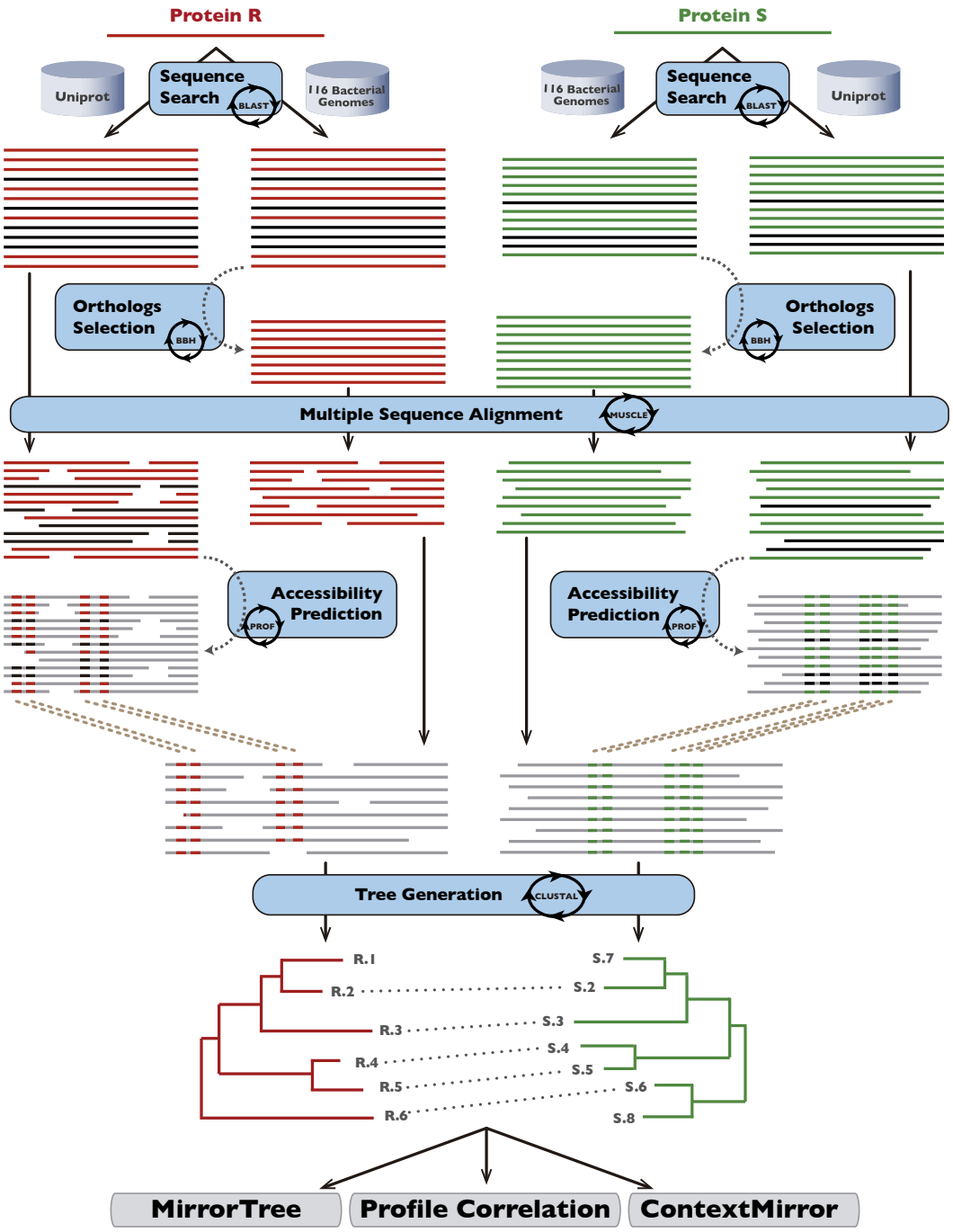


Figure 3.2: **Incorporating predicted solvent accessibility on *mirrortree*-based protein interaction prediction**. This approach tries to evaluate the similarities between the trees of proteins R and S considering only the residues fulfilling a given predicted accessibility criterion. First, we look for protein orthologs in a database of 116 fully sequenced genomes. For each protein, a multiple sequence alignment was generated with these orthologs. In parallel, another multiple sequence alignment was built, this time including every homolog sequence (orthologs and paralogs) found in the Uniprot database. Solvent accessibility for this second alignment was predicted using the PROF program. Based on different criteria, the accessible residues were mapped in the first alignment and the buried residues were excluded for further analysis. The resulting multiple sequence alignments, containing only the positions fulfilling a certain accessibility criteria, were used to model the phylogenetic trees. These trees serve as input for three predictors of protein-protein interactions.

3.3.1 Solvent accessibility prediction

In order to perform a genome-wide accessibility prediction, it is necessary to compile the evolutionary information for each *E. coli* protein. Therefore, a list of candidate homolog sequences was retrieved searching with BLAST [213] in the non-redundant Uniprot database [212]. Sequences with an *E*-value $\geq 1 \times 10^{-4}$ or an identity $< 20\%$ were not considered as potential homologs. Proteins with an alignment coverage lower than 60% (either respect to the hit or the query) were also excluded. Using all the remaining sequences, an MSA was generated for each *E. coli* protein using MUSCLE [215]. Identities with the reference sequence were recalculated considering only those positions in the MSAs with less than 90% of gaps. Sequences with recalculated identities below 20% were also discarded. Additionally, sequence redundancy was removed at 95% to avoid overrepresentation of some sequences on accessibility prediction.

Finally, these MSAs were used as input in the PROF program for predicting solvent accessibility [217]. The resulting scores obtained for every position in the MSA were mapped in the original *E. coli* protein for further analysis. For comparative purposes, an equivalent accessibility prediction was performed with the alignments based on the same MSAs of orthologs used to generate the phylogenetic trees (more details in the next section).

3.3.2 Generating phylogenetic trees

For comparative purposes, we looked for protein orthologs of *E. coli* proteins in a set of 116 fully sequenced organisms previously studied by Juan et al. [164]. In order to avoid sequence redundancy, this set of prokaryotic genomes was designed based on the available genomes at that time, but containing only one strain per species. For each *E. coli* protein, we searched for ortholog sequences using a BLAST "Best Bi-directional hit" (BBH) criterion, with an *E*-value cut-off of 1×10^{-5} and requiring an alignment coverage of 70%. BBH is a standard method to assign orthology to two proteins of two different genomes when they are the best BLAST match of each other in the other genome. The resulting lists of ortholog sequences obtained by BBH were aligned using MUSLCE [215] with the default parameters. Using the previously calculated PROF predictions, we mapped the accessibility information on this MSA of orthologs using the *E. coli* sequence as reference. Different sub-alignments of orthologs were generated including the positions fulfilling the following accessibility criteria:

- eRIA0. Positions predicted as accessible with any value of "reliability".
- eRIA3. Positions predicted as accessible with reliability > 3.
- pACC2. Positions with a predicted solvent accessible surface $> 2\text{Å}^2$.
- pACC12. Positions with a predicted solvent accessible surface $> 12\text{Å}^2$.
- pACC50. Positions with a predicted solvent accessible surface $> 50\text{Å}^2$.
- ALL. No filtering is applied and thus the phylogenetic trees are based on all the positions in the alignment.

For the MSAs in these six different datasets, phylogenetic trees were created using the neighbor-joining algorithm implemented in ClustalW [218], excluding gaps for the distance calculation.

3.3.3 Comparing protein interaction predictions

Pairwise distances between all the orthologs in the each of the phylogenetic trees were calculated for the 6 different sets. These distances were calculated by adding the lengths of the branches separating the corresponding leaves. These distance matrices served as input for three *mirrortree*-based approaches.

Using these distance matrices, the original *mirrortree* (section 3.1.1.1) was applied for protein pairwise comparison in all the accessibility filtered sets. Only the pairs of phylogenetic trees with, at least, 15 species in common and a tabulated *P*-value $\leq 1 \times 10^{-5}$ associated to the correlations were considered. The *mirrortree* correlation coefficients fulfilling the aforementioned conditions were used as input in the Profile correlation methodology (section 3.1.1.2). The same thresholds of 15 species and *P*-value $\leq 1 \times 10^{-5}$ were applied in PC. Finally, we used the ContextMirror algorithm (section 3.1.1.3) applied to the co-evolutionary profiles, to predict protein-interactions using different levels of specificity.

The performance of the three methods when fed with phylogenetic trees generated with residues of different predicted accessibility was evaluated using ROC analysis (section 3.1.3), based on three datasets representing protein interactions of different nature (section 3.1.2).

3.4 Selection of reference organisms

In this section, we investigate the implications of the reference set of organisms used for building the trees in the performance of protein-protein interaction predictions. Nowadays, the number of fully sequenced genomes are orders of magnitude larger than at the times when most of the *mirrortree*-like methods were developed. Therefore, we propose a comprehensive assessment of protein interaction prediction, using different combinations of taxonomic samplings of reference organisms, co-evolution-based methods and types of interaction. For comparative purposes, the 214 Eubacteria and Archaea available in the Integr8 database at the time Juan et al. [164] performed their study were used as initial set. We then sampled this initial set according to different taxonomic criteria using *E.coli K12* as reference organism. Using the samplings, as well as the whole set, we evaluated the performance of three different *mirrortree*-like approaches predicting different types of interactions (Figure 3.3). More details are described in the following sections.

3.4.1 Selection of different subsets of organisms

We mapped the 214 organisms under study using the hierarchy of the NCBI taxonomic tree [219]. This resource classifies the organisms in the public sequence databases in a taxonomic hierarchy, ignoring the quantitative information on phylogenetic distances between them. Since the whole

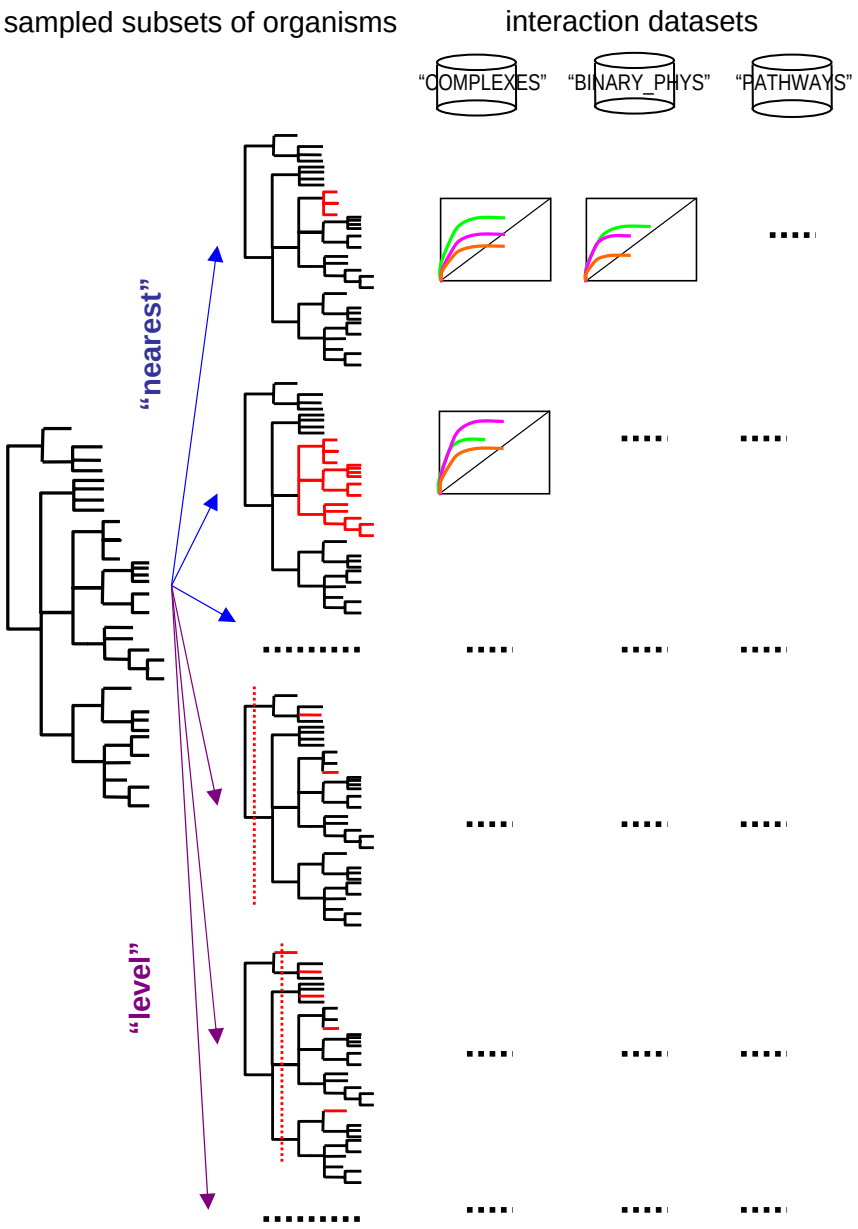


Figure 3.3: **Selection of organisms for protein interaction prediction**. The influence of the reference set of organisms to perform protein interaction prediction is evaluated using sets of organisms taken accordingly with different taxonomic criteria, different evidence of interaction and different methodologies. From an initial set of organisms with completely sequenced genomes (left), a number of subsets (red) are defined based on 2 taxonomic criteria: nearest (blue): going from the reference taxa to the root taking all the organisms in each clade; and level (purple): the tree is successively cut at each taxonomic level and one organism is chosen from each of the resulting groups. Moreover, three different gold standard interaction datasets, based on different types of relationships were used: binary physical interactions, co-membership in the same complex and co-presence in the same metabolic pathways. For each combination of organism selection and interaction dataset, the performance of three *mirrortree*-based methodologies is assessed with a partial-ROC analysis (colored curves).

tree includes information on thousands of prokaryotic taxa, we simplified the tree to represent the taxonomic relationships between the 214 Archaea and Eubacteria of interest. We sampled this subtree, in order to obtain subsets of organisms consistent with the following taxonomic criteria:

- *Nearest*. Starting from *E. coli K12*, we successively take all the organisms belonging to a particular taxa in the path from this reference organism to the root (Figure 3.3). This way, "nearest 1" contains the *E. coli* strains (4 organisms), "nearest 2" includes the Enterobacteriaceae family (21 organisms), and so on, up to "nearest 6", which represents the Bacteria superkingdom (195 organisms); and "nearest 7" the whole dataset (214 organisms). This sampling intends to evaluate the effect of considering species close to the reference organism versus more complete representations of the tree of life. Moreover, this approach assesses the effect of using redundant taxa, such as strains of the same organism, in the results of protein interaction prediction.

- *Level*. This sampling tries to get informative representations of the whole taxonomic tree at different levels of depth. The tree is successively cut at each taxonomic level (superkingdom, phylum,...) of the hierarchy and one organism is selected for each of the taxa defined by that level (Figure 3.3). The criterion for selecting an organism within a taxa is to use that with the highest number of proteins in its genome, in order to optimize the number of possible orthologs in the subsequent steps of the process. As consequence, "level 1" contains 2 organisms (one eubacteria and one archaea) and "level 2" contains 16 organisms, one for each phylum. This way, "level 9" represents the whole dataset. Alternatively to the "nearest" approach, here we present a gradient of sequence redundancies, going from small number of non-redundant sequences in the first level, to the whole redundant dataset in the last level. This design is intended to evaluate the effect of sampling the taxonomy homogeneously at different levels of granularity.

- *Referent set*. For comparative purposes, we included the set of genomes used by Juan et al. [164]. This set, which includes 116 genomes, was obtained selecting only one strain for each species. This set intends to have a representation of the whole tree of life but avoiding the biases introduced by very close genomes in the interaction predictions. The resulting set is very similar to our "level 5" (97 organisms).

In order to assure a reliable estimation, protein interaction prediction usually requires a minimum of 15 organisms in common between the phylogenetic trees under study. For that reason, some of the subsets are never used in practice. The lists of organisms in the final 12 subsets used, as well as a representation of their taxonomic distributions are shown in the Appendix A.

3.4.2 Generating phylogenetic trees

In this study, we varied the set of organisms serving as reference to find ortholog sequences. We generated phylogenetic trees for every *E. coli* protein using the organisms in each of the 12 subsets. In order to do so, we look for protein orthologs in the corresponding list of sampled proteomes using

the "BLAST best bi-directional hit" criterion. Putative orthologs required an *E*-value $\leq 1 \times 10^{-5}$ and an alignment coverage of 70% to be considered for analysis. The set of resulting sequences are then aligned using MUSCLE [215] with the default parameters. The phylogenetic trees are finally created by the neighbor-joining algorithm implemented in ClustalW [218], excluding the gaps for the distance calculations.

3.4.3 Comparing protein interaction predictions

For each phylogenetic tree based on each of the 12 organism sets, a matrix of cophenetic distances is calculated by adding the branch lengths separating the corresponding leaves. These distance matrices served as input for the three co-evolution-based predictors of protein-protein interactions under study.

Mirrortree methodology (MT) (section 3.1.1.1) was applied to the distance matrices obtained using the different selection of organisms. Every pairwise comparison was performed excluding protein pairs with less than 15 organisms in common or a *P*-value $> 1 \times 10^{-5}$. With the matrix of pairwise correlation coefficients fulfilling these thresholds, Profile Correlation (PC) (section 3.1.1.2) was calculated using the same parameters. Finally, the co-evolutionary profiles were compared under the Context-Mirror methodology (CM) (section 3.1.1.3), using a number of different third proteins for partial correlation calculation.

Considering a given subset of organisms, all pairs of proteins in the *E. coli* genome fulfilling the aforementioned requirements were ranked based on the scores of the three methods. We applied ROC analysis (section 3.1.3) to these lists to assess the capacity of the methods to separate positives and negatives on the basis of three different types of interactions (section 3.1.2): binary physical, co-membership in the same macromolecular complex or co-presence in the same metabolic pathway (Figure 3.3). Additionally, we evaluated the ability of the different predictors to recover the maximum number of positives with the best possible accuracy. For the different combinations of datasets, we quantified this tradeoff using the "F-measure" (section 3.1.3). The maximum "F-measure" was used to compare the different predictions assuming that, at this cutoff, the predictor displays its optimal performance.

3.5 Improving the significance of co-evolution detection

In this section, we introduce and evaluate a revisited version of the *P*-values associated to the correlation of distances in *mirrortree*-based approaches. In order to avoid some of the problems affecting the current way of evaluating tree similarity (section 1.7), we designed a new methodology denominated *p-mirrortree*, in which the correlation significance is reassessed by comparing the observed correlation with a null distribution of correlations obtained from the similarities of a large set of permuted phylogenetic trees. Consequently, tree similarities are re-scored using a *P*-value which takes into account the dependencies present in the set of phylogenetic trees under

study. To illustrate the problems associated to the previous versions of *mirrortree* and the improvement obtained by this new method, we compared the original and the new algorithm based on the "historical" sets of organisms available at different time points in the past. Finally, we evaluated the *p-mirrortree* ability to take advantage of the whole matrix of pairwise tree similarities, in a way similar to the PC method (section 3.1.1.2).

3.5.1 *p-mirrortree*

A new methodology denominated p-mirrortree (pMT) to evaluate protein co-evolution was introduced. The key point of this approach is the generation of a null distribution of tree similarities based on the observed distances in a background set of randomized phylogenetic trees.

The background set, which can contain all the phylogenetic trees in an organism or any subset of them, serves as reference to calculate the expected background distribution of tree similarities. Since the correlation coefficients between protein families composed by protein orthologs in many organisms are influenced by the number and characteristics of these organisms, the expected similarity needs to be evaluated in the context of the set of organisms shared by the trees. To build a reliable null model, all the pairwise combinations between phylogenetic trees on the reference set are split into groups based on the number of organisms in common, and a null model is derived for each group independently (Figure 3.4). The size of the groups is defined in a logarithmic scale to add more sensitivity to the correlation changes in trees sharing a low number of leaves. Depending on the total number of organisms used to model the trees and the computational resources available, a smaller or larger number of size groups can be used. For each one of these groups, a iterative process is carried out in order to obtain its corresponding null distribution of tree similarities (Figure 3.4). For each iteration, a pair of trees is randomly sampled with replacement and their distance matrices are retrieved from a pool of pre-calculated matrices of cophenetic distances. A pair of sub-matrices were extracted from the original matrices containing only the distances between sequences belonging to the organisms shared by both trees (Figure 3.4). The resulting sub-matrices are standardized by subtracting from each value their mean and dividing by their standard deviation. Once both matrices are in the same scale, the values corresponding to a given organism are switched between both families. Finally, the distance matrices are de-standardized using their original mean and standard deviation and completed by the distances between organisms present in the original matrices but not shared by the trees. The resulting matrices are returned to the pool in replacement of the original ones and are available for further iterations. Therefore, a single matrix can switch rows/columns multiple times with different matrices. After a number of iterations, the pool contains randomly generated distance matrices but always limited to the distance information available in other trees, reducing the space of possibilities to the ones that have already occurred (Figure 3.4). Finally, for each size group, all posible pairwise correlation coefficients are calculated based on the shuffled matrices, generating a background distribution for that size group.

Once these background distributions are calculated, the significance of a given *mirrortree* correlation coefficient based on a number of organisms in common can be evaluated. This result is

quantified by calculating the probability (*P*-value) of finding a coefficient higher than the observed one in the corresponding background distribution. A low *P*-value indicates a tree similarity much higher than those observed between shuffled trees with similar characteristics and, consequently, is indicative of a meaningful co-evolution.

3.5.2 Generating phylogenetic trees

Using the completely sequenced Eubacteria and Archaea present in the KEGG database (release 59.0 - August 2011) [220], we created phylogenetic trees for all *E. coli* proteins. Prokaryotic protein families with both paralogs and orthologs were retrieved for each protein using KEGG orthology groups. In order to select a single ortholog for each organism, we selected the sequences, which were best ranked against the corresponding *E. Coli* protein on the pre-calculated lists of "BLAST best bi-directional hits" stored in KEGG. The resulting protein families were then aligned by MUSCLE [215] using the default parameters. For each one of the 2,844 MSAs, a phylogenetic tree was created using the neighbor-joining algorithm implemented in TreeBeST [221].

3.5.3 Year-based selection of reference organisms

For each year, from 1995 to 2010, we created two different sets of reference organisms, "redundant" and "non-redundant". The "redundant" list of organisms contains the fully-sequenced Eubacteria and Archaea included in the KEGG database [220] in the corresponding year. A "non-redundant" set was obtained from it by removing the evolutionary close organisms. In order to evaluate which organisms are redundant, pairwise identities between ortholog sequences were calculated using the aforementioned MSAs (section 3.5.2). If two proteomes have more than 70% of the orthologous with 95% or more sequence identity, one of them is excluded. We ran this iterative process starting from the organism with the highest sequence identity with *E. coli* to the one with the lowest. The total number of organisms present in both datasets is shown in Figure 3.5. To assure a minimum number of 15 organisms in the "redundant" and "non-redundant" datasets, we focused on the period 2000–2010 for further analysis.

3.5.4 Year-based distance matrices

For each phylogenetic tree generated based on the available genomes in 2011 (section 3.5.2), a matrix of cophenetic distances was calculated by summing the lengths of the branches separating each pair of proteins. Depending on the size of the protein family, the size of these squared matrices range from 3 rows/columns to the total number of organisms in the corresponding reference set. The distance matrixes for each year were constructed by taking the rows/columns corresponding to the organisms in the redundant and non-redundat sets for that year. For comparative purposes, neither MSAs no trees are recalculated, so distances between a particular pair of sequences in a given protein family remain constant independently of the set of organisms used as reference. As a result, for each protein in the *E. coli* genome, year-based distance matrixes were obtained, based on the redundant or non-redundant sets of organisms available at that time.

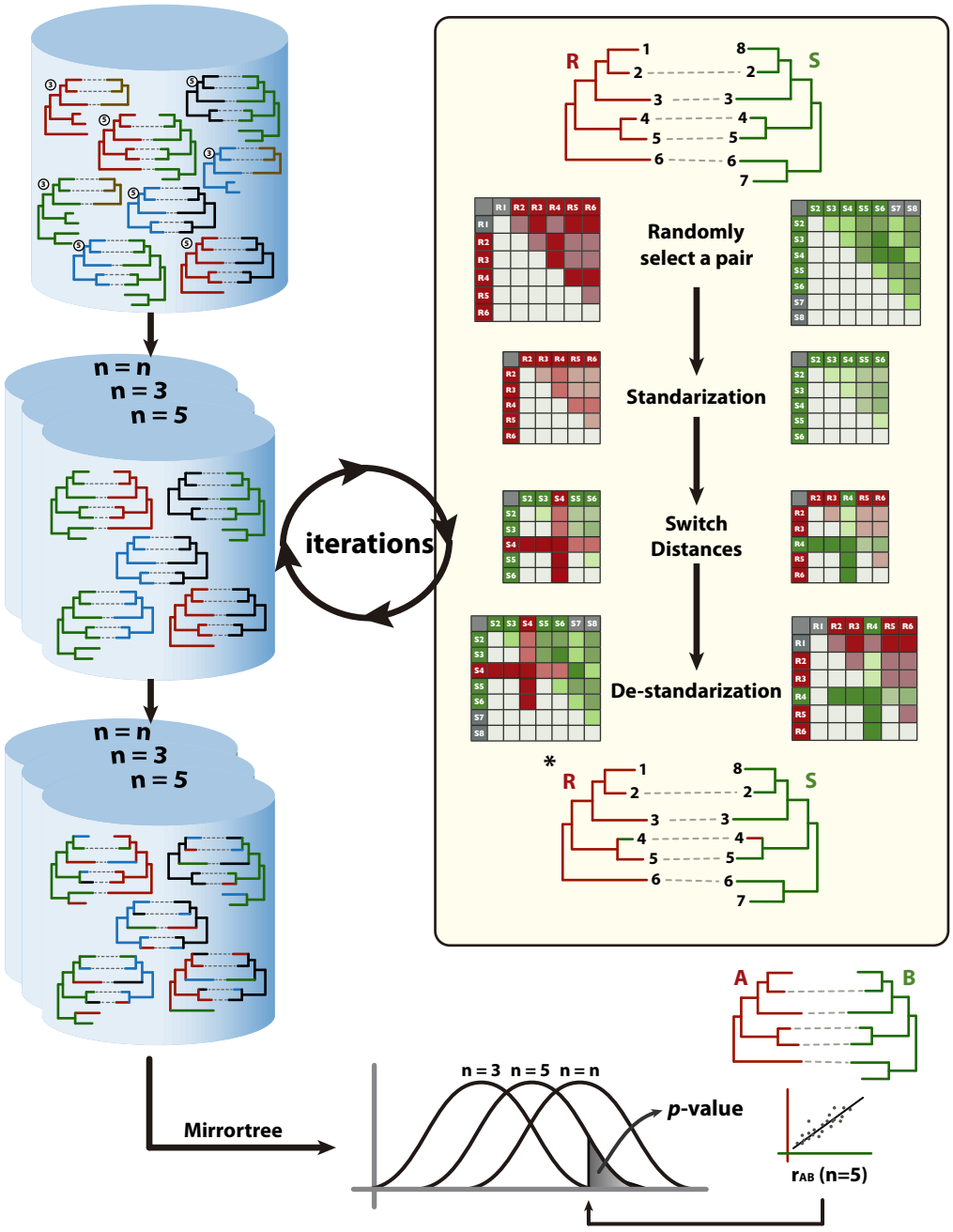


Figure 3.4: **Generating p-mirrortree null distributions**. In the first step all the pairwise combinations between phylogenetic trees are split into groups based on the number of organisms in common - red bubbles. For each group of pairs of trees a number of iterations of a distance swapping procedure are run in order to randomize the trees present in the set. In each iteration, a random pair of trees is selected and standardized based on the distances between sequences belonging to the organisms in common. Rows/columns with the distances belonging to the same organism are swapped between matrixes with a given probability. The resulting matrix are de-standardized to restore their original scales. Both phylogenetic trees are introduced again in the pool of trees for further iterations. The final set of shuffled trees is used to calculate the background distribution of tree similarities. These distributions are used to quantify the statistical significance of an observed tree similarity score. (*) The trees with the swapped branches are shown to illustrate the rationale of the approach, but all the process is applied to the distance matrixes only.

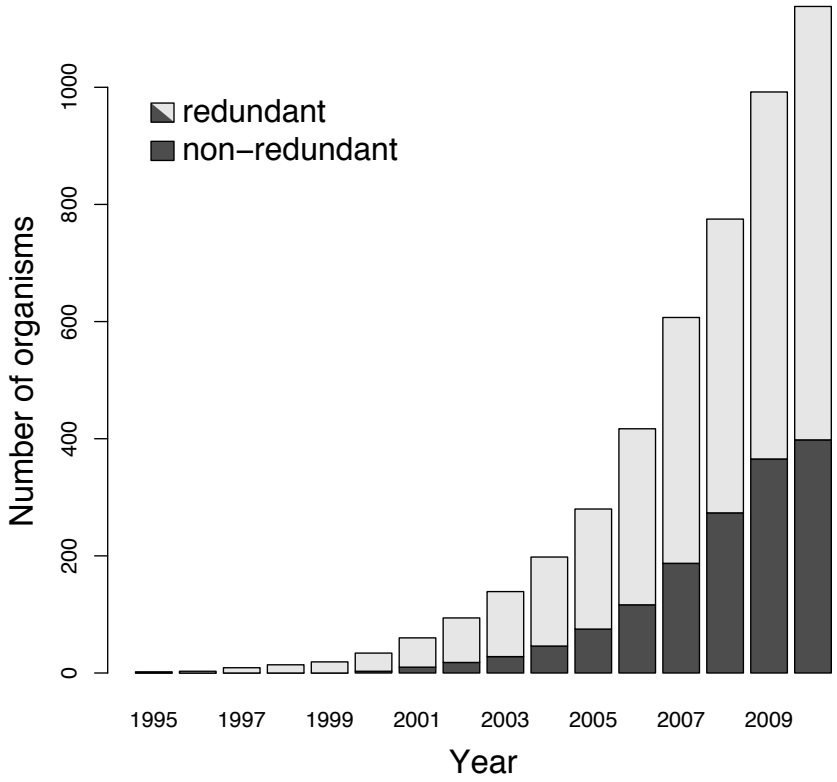


Figure 3.5: **Number of completely sequenced Eubacteria and Archaea available in KEGG for each year in the period 1995–2010.** In dark grey, we represent the number of organisms present in the "non-redundant" set, within the total number of "redundant" organisms represented by the whole bar.

3.5.5 Comparative performance analysis

The original *mirrortree* (section 3.1.1.1) and *p-mirrortree* (section 3.5.1) were applied to each of the "historical" sets of distance matrices (section 3.5.4). The matrices contained the distances between the protein sequences belonging to the available organisms over the years, considering two different redundancy criteria. For the *p-mirrortree*, we executed the algorithm creating a maximum of 40 intervals of number of organisms in common, and ran 1.000 permutation steps with a branch-switch probability of 0.05 (section 3.5.1). Both methods produced lists of putative interacting pairs ranked by their corresponding scores, correlation coefficient or *P*-value, respectively. Those pairs whose proteins are present in the same KO group were excluded to avoid artifacts caused by extremely similar trees.

In order to compare the performances within the same year (redundant/non-redundant and *mirrortree/p-mirrortree*), only those protein pairs present in the four results lists were considered, limiting the evaluation to the discriminant capacity of the scores. The resulting ranked lists were evaluated in the context of ROC analysis (section 3.1.3) using the three types of reference interaction sets previously described (section 3.1.2). Complementary evaluations were performed using only the pairs of trees with at least 15 and 30 organisms in common.

3.5.6 Context-based *p-mirrortree*

The same way the genome-wide matrix of pairwise *mirrortree* correlation coefficients is used by the PC method (section 3.1.1.2) to improve the prediction of interactions, we evaluated how a similar approach works with this new score (*p-mirrortree P*-value). Using the organisms dataset of 2010, we applied the PC method to the matrix of pairwise *mirrortree* scores, and a similar approach to the matrix with the new *P*-values (Figure 3.6). We evaluated both predictions based on the "Complexes" gold standard (section 3.1.2) using ROC analysis (section 3.1.3).

Additionally we introduced a method named Hierarchical Co-evolutionary Analysis (HCA) to explore the "co-evolutionary hierarchy" defined by the *P*-values. The distance between a pair of *p-mirrortree* coevolutionary profiles was defined as 1 minus the PC score previously defined. Using these distances, different clustering algorithms were applied. These include Ward's minimum variance, complete linkage, neighbor-joining and UPGMA. This generates a hierarchy of co-evolutionary relationships between *E. coli* proteins (Figure 3.6). The accuracies of the top-scoring results were calculated and compared with those obtained from PC using either correlation coefficients or *P*-values as input. Additionally, the biological meaning of some co-evolutionary groups showing up in this clustering is evaluated.

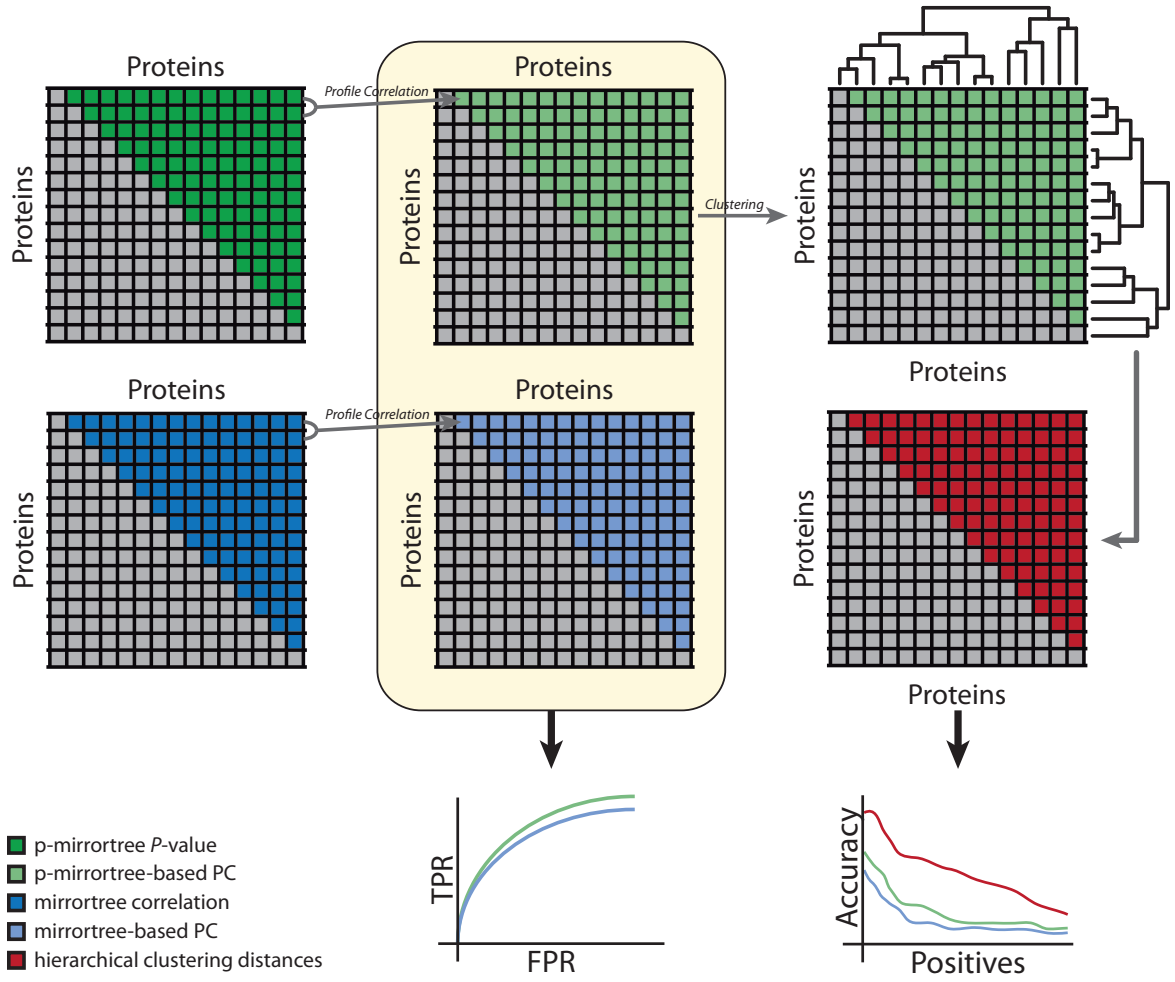


Figure 3.6: **Context methods based on mirrortree and p-mirrortree results**. Profile correlation were calculated using both genome-wide MT correlation coefficients (dark blue) and pMT *P*-values (dark green). The results of both methodologies (light blue and light green, respectively) were evaluated using the "Complexes" set (section 3.1.2) as gold standard. The True Positive Rates (TPR) and False Positive Rates (FPR) were drawn for the different possible thresholds in the context of ROC analysis (section 3.1.3). Moreover, a hierarchical clustering was applied to pMT PC results and the cophenetic distances of the resulting clustering were used as a new scoring schema. The performance of this score predicting interactions in the "Complexes" set (section 3.1.2) was evaluated within the top ranked results and compared with that of PC applied over MT correlations or pMT *P*-values.

Results

4.1 Mirrortree web server

The Mirrortree Server (http://csbg.cnb.csic.es/mtserver) enormously simplifies the pairwise comparison of phylogenetic trees, allowing non-expert users to interactively study the co-evolutionary features of a given pair of families using, in the simplest case, single representative sequences as input. On the server side, the process for generating the trees and preparing the data and results for the client takes around 10 minutes (800 residues and 120 species in common).

When the aforementioned workflow (described in detail in section 3.2) is completed at the server side, the user receives an e-mail containing a link to the results. There, together with several intermediate files such as MSAs, phylogenetic trees or statical graphical representations of the mirrored trees, the user finds a flash-based application to interactively explore the mirrored trees (Figure 4.1). This tool is visually dominated by the representation of both phylogenetic trees connected by lines linking the sequences belonging to the same organisms.

To confront paired clades and improve the representation, tree branches can be swapped by "drag and drop" or using the "Swap" button. The user can also zoom and scroll over the canvas selecting different leaves or clades to restrict the co-evolutionary analysis to certain groups of organisms.

Floating over the tree representation, different panels provide complementary tools that can be freely moved, resized or hidden in a window-based interface. Among these panels, that labelled as "MirrorTree Results" shows in a color scale the tree-tree similarity calculated with the standard *mirrortree* algorithm (Figure 4.1). The similarity for the current selection of orthologs (in red in the tree representation) is also shown in this panel. Both results display the tabulated *P*-value associated to their correlation coefficient, considering the number of organisms in common.

By clicking the leaf labels (protein IDs), additional information is retrieved into the "Protein Info" panel (Figure 4.1), including protein name and FASTA sequence from Uniprot [212] or previously reported interactions from IntAct [81]. Alternatively to the direct selection of organisms, tree clades can be selected based on taxonomic criteria using the "NCBI taxonomy tree" panel (Figure 4.1). In this panel, a hierarchical representation of the NCBI taxonomy tree [219] is displayed using only those organisms present in the current version of the Integr8 database. This

tool enormously simplifies the evaluation of co-evolutionary events related with speciation events in certain kingdoms or families. Moreover, using the “Export selection” button, the user can export the selected sequences as MSA or in raw format, in order to perform further analysis.

Finally, a scatter plot with a simplified representation of the correlation between inter-protein distances in both families is also shown (Figure 4.1). Switching the combo box in the panel, the user can easily change the representation to affect all distances, or only those involving the selected organisms. This panel is particularly useful to look for clouds of points far from the general trend of distances (i.e “outliers”). Those distances, decreasing the overall similarity of the trees, can indicate evolutionary trends restricted to certain groups of organisms and might be related to non-standard evolutionary events such as horizontal gene transfer [132]. Selection of points in the plot implies the selection of corresponding sequences in the trees.

All these features, as well as some other characteristics of the server, are extensively explained in the “Help” web page of the server. Users can also find a detailed tutorial illustrating the kind of analysis that can be performed in the server.

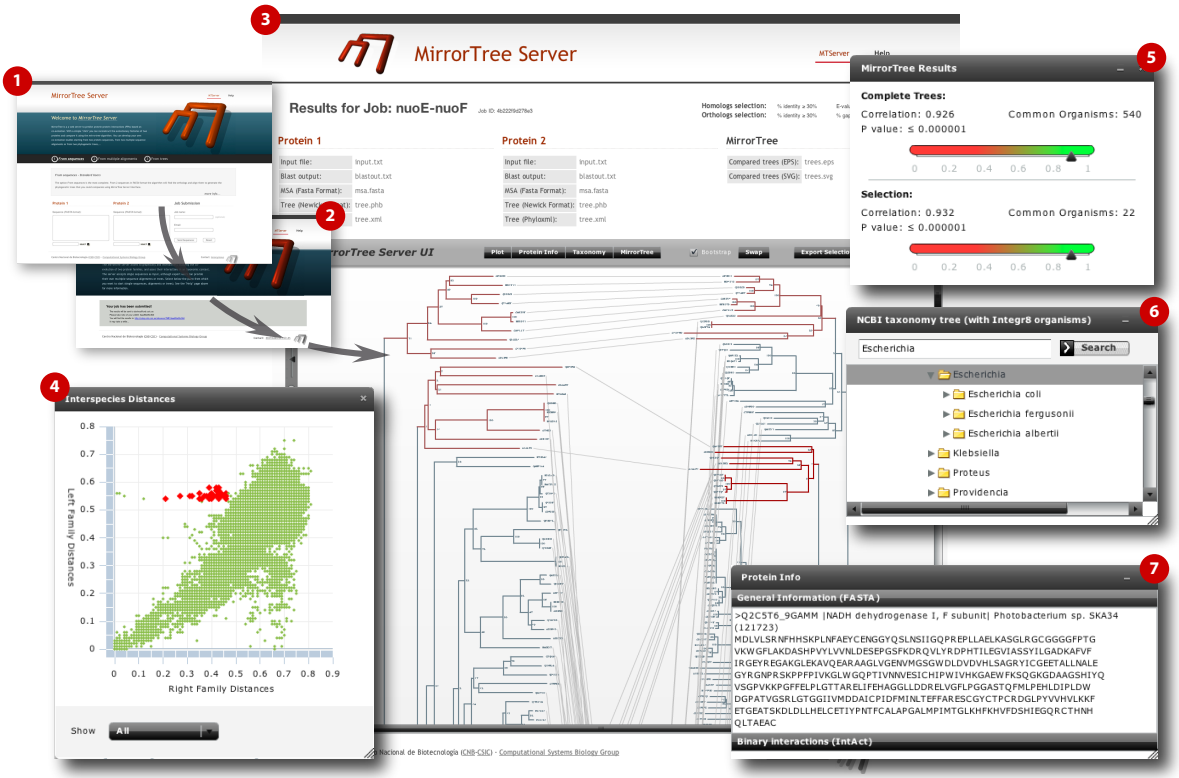


Figure 4.1: **Mirrortree Server User Interface**. **(1)** Job submission page. **(2)**. Job status. **(3)**.Results page containing intermediate files and the Mirrortree Server User Interface. The floating panels can be shown/hidden and freely moved/resized as independent windows. The Mirrortree Server User Interface can also be maximized to be used in full-screen mode. **(4)**. Distance correlation plot panel. **(5)**. Tree and sub-tree correlation coefficients, number of organisms in common and *P*-value. **(6)**. Taxonomy browser. **(7)**. Additional protein information retrieved from Uniprot [212] and IntAct [81].

4.2 Incorporating information on predicted solvent accessibility

The impact of incorporating predicted solvent accessibility on protein interaction prediction based on similarity of phylogenetic trees was evaluated using different setups, as described in the corresponding methodological section (section 3.3). Figure 4.2 shows the performance of the three *mirrortree*-based approaches, when using the phylogenetic trees generated from residues of different predicted accessibility, and evaluated based on the three different datasets of protein interactions. Similarly, Figure S1 shows the same results with different scales for each plot, in order to highlight the differences within the same predictor and dataset of interactions.

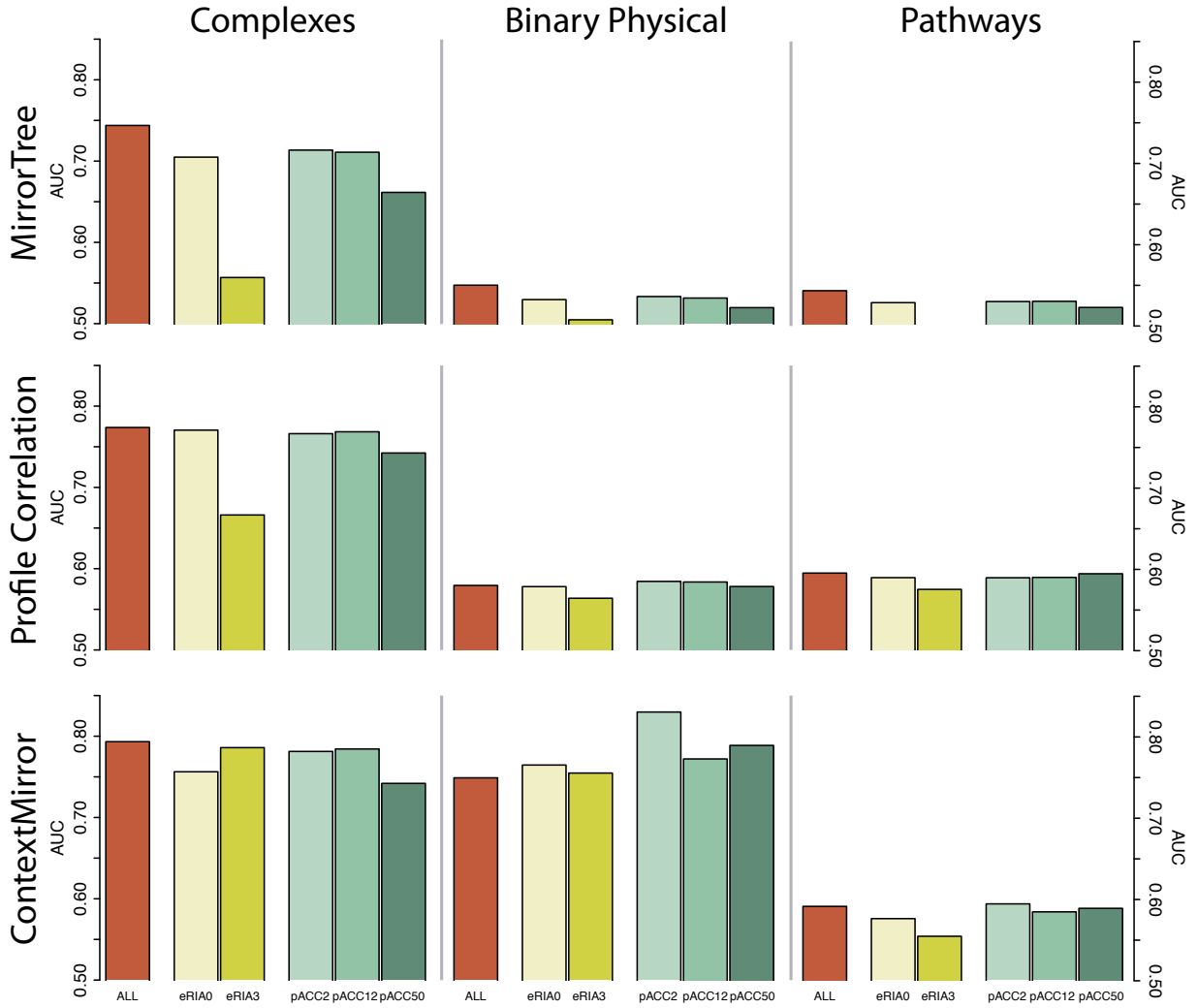


Figure 4.2: **Performances for different combinations of prediction methods (rows), interaction gold standard (columns) and predicted accessibility filter (colors)**. Performances are evaluated in terms of Area Under ROC Curve (AUC). The same figure with different scales for each plot is available as Figure S1. Equivalent figures with the results obtained using predicted accessibility derived from MSAs of orthologs are available as Figure S2 and Figure S3.

In agreement with previous results [210], the method's performances suggest that protein interactions are more easily detected when direct or indirect physical relationships are established between the interacting partners. Physical interactions, especially those between proteins in the same complex, are predicted with better performances than functional interactions taking place between proteins in the same metabolic pathway. Regarding the methods themselves, recent approaches such as PC and CM perform better than the original MT, which presents proper performances only when detecting interactions between proteins on the same macromolecular complexes.

Regarding the impact of adding predicted solvent accessibility information, in most cases the removal of non-accessible residues worsens the results (section 3.3.2 and Figure S1). However, for the case of binary physical interactions, the results of the PC and CM are improved when using only residues predicted as accessible for constructing the trees. The highest improvement is obtained when those residues with an area predicted as accessible larger than 2Å^2 (pACC2) are considered. Although restricting the prediction to the residues predicted as exposed (eRIA0), or with a predicted accessible area $> 12\text{Å}^2$ (pACC12) or $> 50\text{Å}^2$ (pACC50), also improves the predictions, they perform worse than the pACC2 set residues.

Under the hypothesis that the co-evolutionary signal is spread through the whole sequence (section 1.6.4.1), residue removal would imply a loss in co-evolutionary information. Consequently, we expect that predictions derived from filtered alignments will perform worse than those based on full sequences. Attending to Figure 4.3, we observe how the performance of MT methodologies in detecting binary physical interactions increases linearly with the logarithm of the average number of positions in the filtered MSAs. In MT, far from helping the prediction of binary interactions, the performance approaches random when the average number of positions drops to 50 residues. This trend has been observed also in MT predictions when evaluated using other types of interactions (Figure S4). Apparently, this is not the case for the binary physical interactions predicted by PC or CM, in which the pACC2 residues represent a proper tradeoff between the enrichment in accessible residues and the global loss of positions. Notice that the linear trend for PC and CM is broken (Figure 4.3), since pACC2 renders the best performances in spite of not having the largest number of positions.

Additionally, we evaluated the effect of adding accessibility information in the method's performance at the top of the ranked lists of protein pairs. In many cases, more than the global discriminative capacity represented by AUC, the users of this methodology may be interested in just a few candidates (the top scoring ones) ignoring the protein pairs at the bottom of the ranked list. Therefore, a real improvement needs to enhance the top scoring pairs and not only the global performance of the predictor. In Figure S5, we assess the top pairs for the same combinations of predictive method, interaction evidence and accessibility filter. Here, we evaluated the predictions in terms of the number of true positive interactions among the top "n" pairs, "n" being the total number of positives of the methods. In a perfect predictor, the number of true positives among these pairs should be equal to "n" (all the positives at the top of the list). CM comes up as the methodology with the lowest number of false positives, as previously reported. The observed benefit of including accessibility information when predicting binary physical interactions extends

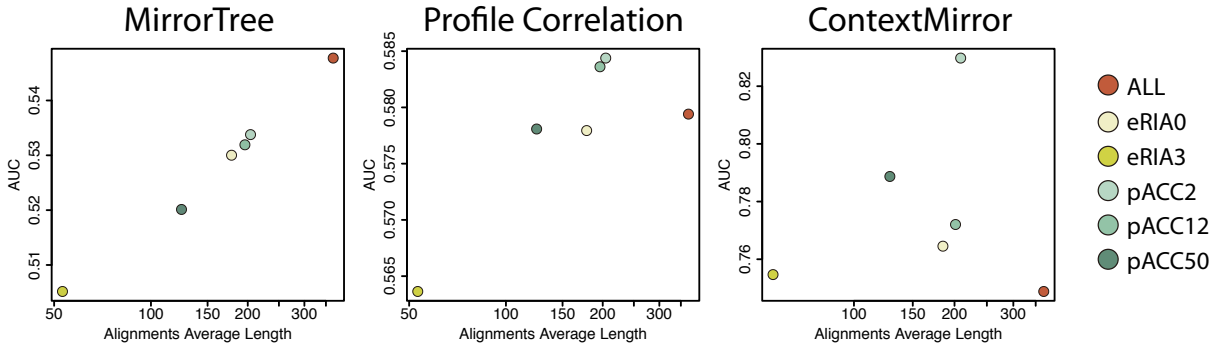


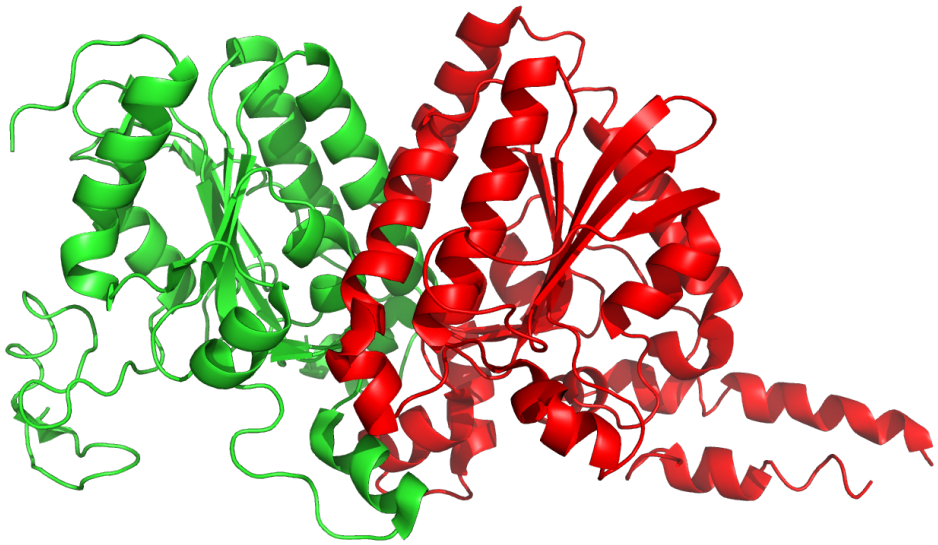
Figure 4.3: **Relationship between the performances of the methods and the average length of the alignments filtered according with different accessibility criteria**. Performances were evaluated in terms of Area Under ROC Curves (AUC). The length of the filtered alignment is the number of positions (fulfilling a given predicted accessibility criteria -colors) used to construct the trees. These results were validated using the dataset of binary physical interactions. The corresponding plots for the other interaction datasets are available as Figure S4.

to the top scoring pairs. Using the pACC2 and the pACC12 sets of residues, the number of true positives in the top "n" results increases, even though the total number of significant positives is smaller than in the set without filtering accessibility.

At this stage, we showed the results of incorporating accessibility information predicted using MSAs based on homolog sequences. Nevertheless, in the framework of *mirrortree*-related methodologies, the step of predicting accessibility could enormously be simplified by using the same MSAs of orthologs used to build the phylogenetic trees. Therefore, we incorporated in this pipeline the accessibility information predicted from the same alignments used to run *mirrortree*-based methods producing Figure S2 and Figure S3 as equivalent figures of Figure 4.2 and Figure S1, respectively. The general drop in performance previously observed when incorporating predicted accessibility looks shaper here. Even in those cases where accessibility information improved the results, the improvement looks significantly lower. As a result, phylogenetic trees obtained from MSAs constructed *ad hoc* from protein homologs prove to be a better option for the incorporation of accessibility information on protein interaction prediction.

4.2.1 Example

To exemplify the effect of including predicted solvent accessibility information on the results of co-evolution based methods, we show a pair of interacting proteins whose co-evolution is more evident when evaluated using accessible residues. Figure 4.4 illustrates the scores of the three *mirrortree*-based methods for the α and β subunits of the *E. coli* acetyl-CoA carboxylase carboxyl transferase using different filters of accessibility. Whereas the MT correlation coefficients get their maximum value when using the whole sequences as input, the PC and CM scores benefit from restricting the analysis to accessible residues. In CM, for instance, the score increases from 0.6068 to 0.6818 by restricting the analysis to the pACC2 residues. Correlation coefficients are also improved using eRIA0 and pACC12.



	ALL	eRIA0	eRIA3	pACC2	pACC12	pACC50
MirrorTree	0.9083	0.8982	0.8331	0.8998	0.9069	0.8924
Profile Correlation	0.9516	0.9531	0.9435	0.9605	0.9592	0.9328
ContextMirror	0.6068	0.6650	0.5410	0.6818	0.6828	0.5407

Figure 4.4: **Example illustrating the effect of incorporating predicted solvent accessibility on the evaluation of tree similarity.** The structure of the complex between the α and β chains of *E. coli* acetyl-CoA carboxylase carboxyl transferase is shown in ribbon representation. The table contains the scores of the three methods for this interacting pair of proteins based on the trees generated using the six different criteria of predicted accessibility.

4.3 Selection of organisms

The protocol followed to evaluate the effect of the set of organisms used for constructing the phylogenetic trees on the performance of co-evolution-based methods is described in the methodologies section (section 3.4).

For each combination of method, interaction dataset and subset of organisms based on taxonomic criteria, we obtained a partial ROC curve which represents the ability of a given method to discriminate interacting from non-interacting proteins. Figure 4.5 shows in different colors the ROC curves obtained using different sets of organisms grouped by the interaction dataset and method. Detailed information on the organisms contained in each dataset is available in Appendix A. Since ContextMirror (CM) produced different lists of predictions as a consequence of its methodological particularities, we show the results for level 10, which previously displayed a good performance in protein interaction prediction [164]. Although the different CM levels vary in terms of accuracy and coverage, their relative performance respect to the sets of organisms are similar, so the rest of the levels were not included for the sake of clarity. In order to highlight the differences between the different plots, Figure S6 shows the same partial ROC curves, but using the same scale.

In parallel, for the same combinations of predictive method, gold standard dataset of interacting proteins and reference set of organisms, the results were evaluated in terms of "F-measure *vs* method score" (Figure 4.6). The optimal compromise between precision and recall is obtained at those cutoffs in which the F-measure presented the highest values. The plots are in the same scale in order to facilitate the interpretation of the results. Since the "F-measure *vs.* score" plot shows the performance of the method at the optimal cutoff, we used the highest F-measure as a single numerical estimator of the predictive capacity. In Figure S7, we compared this maximum value for the different sets of organisms in each of the interaction datasets and predictive methods.

Independently of the approach used for evaluation, the results indicate that all these co-evolution based methodologies are able to predict protein interactions of different nature across a wide range of organism sets (Figure 4.5, Figure 4.6 and Figure S7). These results are in line with the growing evidence that protein interactions are closely related with co-evolution events. Many groups using different variations of the methodologies fed with trees based on diverse sets of organisms and independent datasets converged to the same conclusion. So it was expected that, in general, the reference set of organisms would not be a limiting factor in the co-evolutionary analysis. However, we can observe that the organisms used for building the phylogenetic trees directly influence the global performance of the methodologies This phenomenon invites to look for optimal sets of organisms depending on the type of interaction we are interested in.

The most obvious results are that, in general, the performance of the methods has a dependence on the set of organisms used, and that not always using more organisms is better. Some approaches such as the Profile Correlation methodology (PC) has proven to be robust in predicting protein interactions, independently of the set of organisms under study (Figure 4.5, Figure 4.6 and Figure S7: panels b, e and h). Only in those cases with a small number of organisms, the

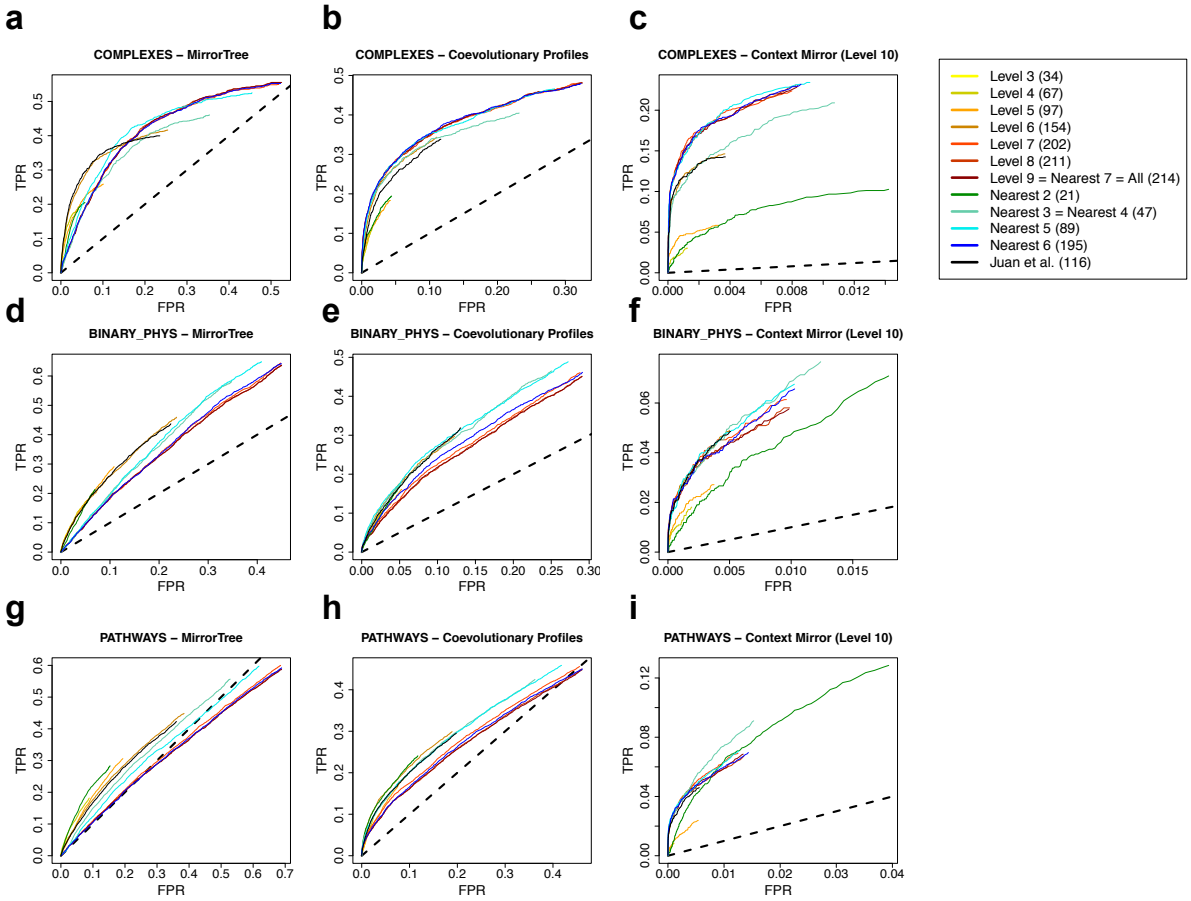


Figure 4.5: **Matrix of partial ROC curves**. The partial ROC curves evaluate a given list of predictions obtained by the combination of a methodology (columns) with a set of organisms (colors according to the legend), evaluated using a given dataset of interactions (rows). In the legend, the number of organisms present in each dataset is included within brackets. The list of organisms in each dataset, as well as a representation of their taxonomic distribution is available in Appendix A. The dashed line represents the performance of a random classifier. The plot scales are adapted to improve the visual comparison of the partial ROCs among the different sets of organisms. Equivalent plots with the same scale are shown in Figure S6

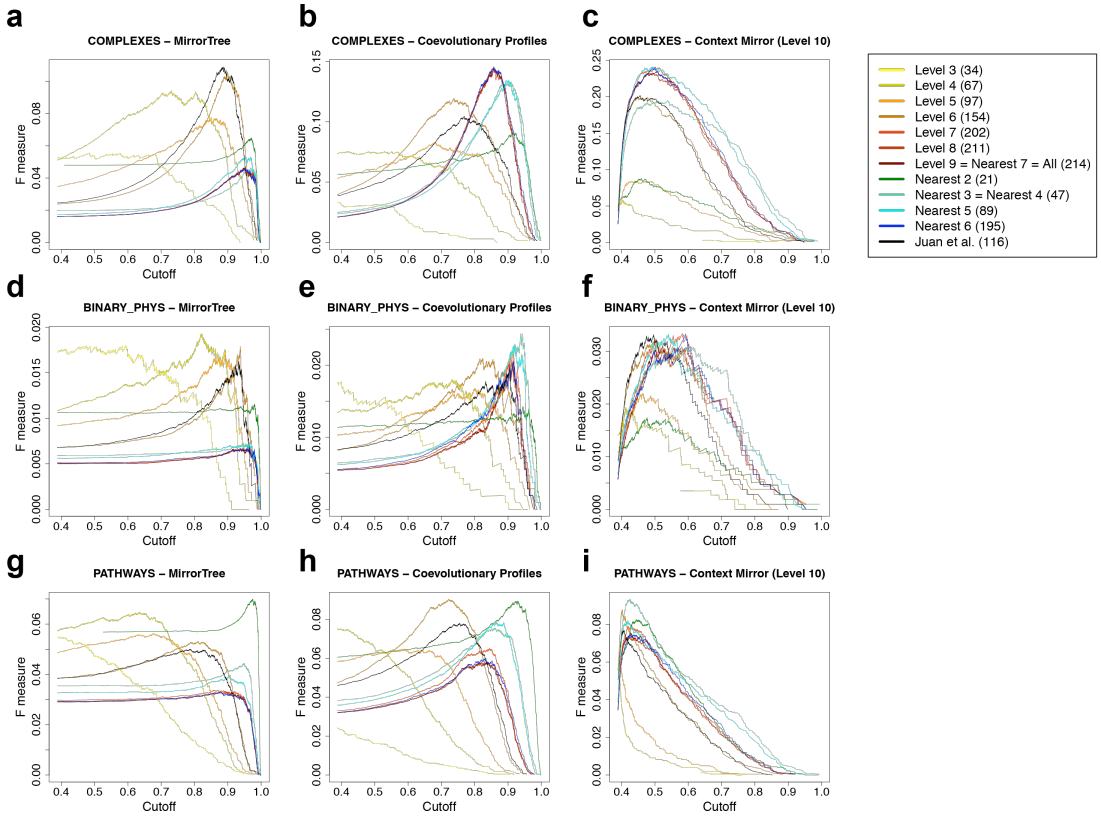


Figure 4.6: **Same results as in Figure 4.5 represented in terms of F-measure vs. score.**

methodology suffers from the lack of information. This relative independence of the organism set is probably due to its ability to filter artifactual tree-correlations such as those related to phylogenetic bias. On the other hand, as previously reported, CM presents the highest accuracies, but at the expense of producing fewer significant predictions (Figure 4.5, Figure 4.6 and Figure S7: panels c, f and h). The global performance of CM seems to drastically drop when the number of reference organisms is reduced. This effect can be easily explained by the fact that CM requires a rich network of genome-wide inter-protein similarities in order to calculate the partial correlations. As the number of organisms decreases, the number of pairwise similarities with less than 15 organisms in common or non-significant increases, making the network of similarities sparser and less usable for CM. Mirrortree (MT) in the other side, achieves the worst global performances (Figure 4.5, Figure 4.6 and Figure S7: panels a, d and c). Moreover, MT is severely affected by the presence of redundant organisms. Whereas PC and CM in general benefit from using as many available organisms as possible ("nearest 6", "nearest 7" = "level 9" = "All"), for MT, this benefit reaches a point where it enters into conflict with the presence of redundant organisms. Consequently, sets of organisms such as "level 6" or "Juan et al." predict more accurately interactions using MT, as they represent the whole tree of life having removed redundant organisms.

Regarding the type of interactions under study, the relationships of co-presence in macromolecular complexes are predicted better independently of the method used, followed by binary physical interactions and co-presence of metabolic pathways (Figure 4.5, Figure 4.6 and Figure S7). Apart from this general trend, each type of interaction appears to be better predicted by certain sets of organisms. In general, complexes are more accurately predicted with trees which include distant organisms, while binary interactions and pathways are predicted better using trees restricted to close organisms such as "nearest 5", "nearest 6" or "nearest 7" when using context-based methods (Figure 4.5).

4.3.1 Examples

In order to illustrate how using close/distant organisms can drastically affect the interaction prediction, we included some examples of proteins extracted from the "Complexes" and "Binary physical" datasets as possible cases of "old" and "recent" interactions, respectively (Table 4.1). The selected proteins involved in the "recent" interactions are mostly related with metabolic functions, except for MinE-MinD, which is associated with the division machinery in certain organisms [222]. On the other hand, the "old" interactions include some proteins involved in transcription/translation machineries as well as interactions between members of a very ancient family of proteins: the ABC transporters [223].

Table 4.2 shows the results obtained for these proteins using PC based on two sets of organisms: "level 9", which contains the full set of organisms; and "nearest 2", which contains only organisms belonging to the Enterobacteriaceae family. For each of these sets and considering a given protein, the total number of candidate interacting partners for which it was possible to calculate significant correlations (Tot) was indicated. Within these candidates, the number of positive interactions (+) was also shown in the same column. In another column, the PC score with the annotated interactor

Table 4.1: Examples of proteins potentially involved in "recent" and "old" interactions

	Protein	Description
	MINE_ECOLI	Cell division topological specificity factor
	MIND_ECOLI	Septum site-determining protein MinD
	PABA_ECOLI	Para-aminobenzoate synthase glutamine amidotransferase component II
recent	PABB_ECOLI	Para-aminobenzoate synthase component 1
Binary physical	DHAS_ECOLI	Aspartate-semialdehyde dehydrogenase
	DNAK_ECOLI	Chaperone protein DnaK
	GSHB_ECOLI	Glutathione synthetase
	AMPM_ECOLI	Methionine aminopeptidase
	DPO3A_ECOLI	DNA polymerase III subunit alpha
	DPO3E_ECOLI	DNA polymerase III subunit epsilon
	RPOB_ECOLI	DNA-directed RNA polymerase subunit beta
old	RPOA_ECOLI	DNA-directed RNA polymerase subunit alpha
Complexes	ZNUB_ECOLI	High-affinity zinc uptake system membrane protein ZnuB
	ZNUC_ECOLI	Zinc import ATP-binding protein ZnuC
	ZNUA_ECOLI	High-affinity zinc uptake system protein ZnuA

is shown (corr), displaying only the highest score (max) when there is more than one positive partner is predicted. Finally, the Area Under ROC Curve (AUC) (section 3.1.3) was also included as a single numerical estimator indicating the predictor ability to distinguish positive and negative pairs for that particular protein along the list of candidate partners. For simplicity the AUC values where calculated using the positives and negatives on the lists, contrary to the partial ROC curves calculated in Figure 4.5, which were based on the total number of positives and negatives in the gold standard dataset.

Table 4.2: PC results for the proteins in Table 4.1 based on two sets of organisms

	Protein	Level9 (=all)			Nearest2		
		Tot/+	AUC	Interactor (corr)	Tot/+	AUC	Interactor (corr)
	MINE_ECOLI	846/1	0.12	MIND_ECOLI (0.52)	223/1	**0.83**	**MIND_ECOLI (0.60)**
recent	PABA_ECOLI	671/1	0.28	PABB_ECOLI (0.49)	106/1	**0.96**	**PABB_ECOLI (0.96)**
	DHAS_ECOLI	760/1	0.17	DNAK_ECOLI (0.48)	384/1	**0.81**	**DNAK_ECOLI (0.90)**
	GSHB_ECOLI	755/1	0.30	AMPM_ECOLI (0.61)	375/1	**0.93**	**AMPM_ECOLI (0.95)**
	DPO3A_ECOLI	306/1	**0.70**	**DPO3E_ECOLI (0.73)**	128/1	0.11	DPO3E_ECOLI (0.57)
	DPO3E_ECOLI	357/1	**0.64**	**DPO3A_ECOLI (0.73)**	123/1	0.22	DPO3A_ECOLI (0.57)
	RPOB_ECOLI	280/7	**0.82**	**(0.98) max**	126/4	0.48	(0.93) max
old	RPOA_ECOLI	258/6	**0.81**	**(0.80) max**	90/3	0.48	(0.93) max
	ZNUB_ECOLI	370/2	**1.00**	**(0.87) max**	129/1	0.36	ZNUC_ECOLI (0.74)
	ZNUC_ECOLI	386/2	**0.99**	**(0.87) max**	123/2	0.41	(0.74) max
	ZNUA_ECOLI	395/2	**0.98**	**(0.87) max**	39/1	0.79	ZNUC_ECOLI (0.74)

On the selected examples, the "recent" interactions present higher co-evolutionary scores and AUCs - in bold - when using the "nearest 2" dataset. Exactly the opposite occurs for the "old" interactions, which show higher co-evolutionary scores and performances when using the organisms present in "level 9". For example, the subunit alpha of the "ancient" machine DNA polymerase III (DPO3A_ECOLI) has only one reported interaction in the "Complexes" dataset which is the epsilon

subunit (DPO3E_ECOLI). If we attend to the results of the co-evolution analysis using both sets of organisms, we can appreciate significant differences. Whereas PC using “level 9” is able to produce significant scores for 306 pairs involving the alpha subunit, this number is considerably reduced to 128 when the “nearest 2” set is introduced. Moreover, the PC score calculated with the full set of organisms as reference (“level 9”) is 0.73, dropping to 0.57 when using only enterobacterias (“nearest 2”). As a direct consequence of the correlation drop, the proportion of false positives grows, negatively affecting the AUC, which decreases from 0.70 in “level 9” to 0.11 in “nearest 2”. We observe exactly the opposite behavior in the “recent” interactions where the “nearest 2” dataset performs better predicting interacting partners (Table 4.2).

4.4 Improving the detection of significant co-evolution

4.4.1 More insight on *p-mirrortree* null distributions

Previous versions of *mirrotree* either disregard the *P*-values associated to the correlation score or use tabulated ones, calculated analytically or derived from random sets of numbers not fulfilling the properties of tree-based distances. In order to get insight into the variations introduced by pMT and their *P*-values specifically derived for the genomic tree comparison problem, we compared the null distributions obtained by both approaches.

The pMT method introduces an additional step in the prediction of protein interactions: the calculation of the null distribution of tree similarities generated as a consequence of the permutation of a “background” set of trees. This process, described in section 3.5.1, creates a number of distributions of expected tree similarities for different intervals of number of organisms in common between the pairs of trees. Some of these distributions for the genomes available at different years are compared with the corresponding distributions of correlations between sets of random numbers (those tabulated and used in previous versions of mirrortree)(Figure 4.7). As expected, the average correlation coefficient of random numbers is always 0. Moreover, as soon as the sets of numbers being correlated are larger, the probability of obtaining “extreme correlations”, either positive or negative, decreases. However, some of these general observations are not extended to the correlation coefficients calculated using distance matrixes of permuted trees. As described in different studies, phylogenetic trees tend to share a background similarity and, consequently, the distribution of correlations is always shifted to higher values. Besides, the expected correlation distributions highly depend on the number of organisms shared by a pair of trees. Figure 4.7 shows that pairs of trees sharing a small set of organisms present a wider range of correlation coefficients, whereas pairs of trees sharing many orthologs present correlations in a narrower range. Contrary to the random numbers, these distributions are not centered in the same value, so the null distribution of correlation coefficients varies depending on the number of organisms in common. Therefore, the probability of obtaining a given correlation coefficient varies depending on the distribution of organisms used to generate the trees.

Here we used the number of organisms in common between a pair of phylogenetic trees in a given set of organisms as a proxy for the set of organisms under comparison. Since the *mirrortree*-based approaches are usually focused on predicting protein interactions in a reference organism, it is expected that the number of organisms in common reflects the conservation of these families along the tree of life. However, this number is highly dependent on the total number of organisms used to generate the phylogenetic trees. Indeed, we need to be aware that a given correlation coefficient calculated with 10 organisms in common has a completely different meaning if the phylogenetic trees are generated using the 38 "non-redundant" organisms available in 2001 or the 335 of 2010 (Figure 4.7). This condition, previously ignored by the tabulated *P*-values, is addressed by pMT, which independently constructs a null distribution for each number of possible organisms in common between the pairs of phylogenetic trees.

4.4.2 Historical assesment of p-mirrortree predictions

We compared the performance of p-mirrortree (pMT) with that of the original mirrortree (MT) in predicting different types of interactions using the set of organisms available in the period 2000–2010, as well as non-redundant versions of these sets (section 3.5).

The performance of MT and pMT predicting two different types of physical interactions are shown in Figure 4.8. The clearest observation is that both methods based on similarities of phylogenetic trees are able to capture part of the co-evolutionary signal related to physical protein interactions. The general trends observed suggest that protein interaction predictions have benefited from the increase in the number of sequenced genomes during the last decade. Indeed, there is no clear evidence suggesting these trends have reach a plateau, so further improvement can be expected over the next few years. Similarly to previous studies, interactions defined as co-membership in the same macromolecular complex present the highest AUCs, followed by binary physical interactions. Interactions based on co-membership in the same metabolic pathway present poor and constant AUCs (Figure S8), suggesting that co-evolution may not be a generalized process between the proteins of the same pathways.

The performance of pMT when predicting physical interactions using the organisms available during the last 10 years is higher than that of MT ($\sim$0.10–0.15 increase of AUC) (Figure 4.8). This improvement is obtained at no cost in terms of applicability, since no-additional restrictions are required to run pMT apart from the contextual information necessary to generate the null distributions.

Previous results suggested that MT predictions are negatively affected by redundant taxa (section 4.3). To obtain further insight into this, we compared MT and pMT performances using the aforementioned year-based lists of organisms against a subset of the same organisms from which redundancy was removed. In Figure 4.8, we observe that MT performances benefit from the usage of the "non-redundant" sets, supporting previous evidences. Indeed, the gap in AUC between "redundant" and "non-redundant" sets becomes bigger as the taxonomical redundancy increases over the years. On the other hand, pMT presents a higher robustness to redundancy. Although "non-redundant" sets achieve slightly better performances, this improvement seems to be constant

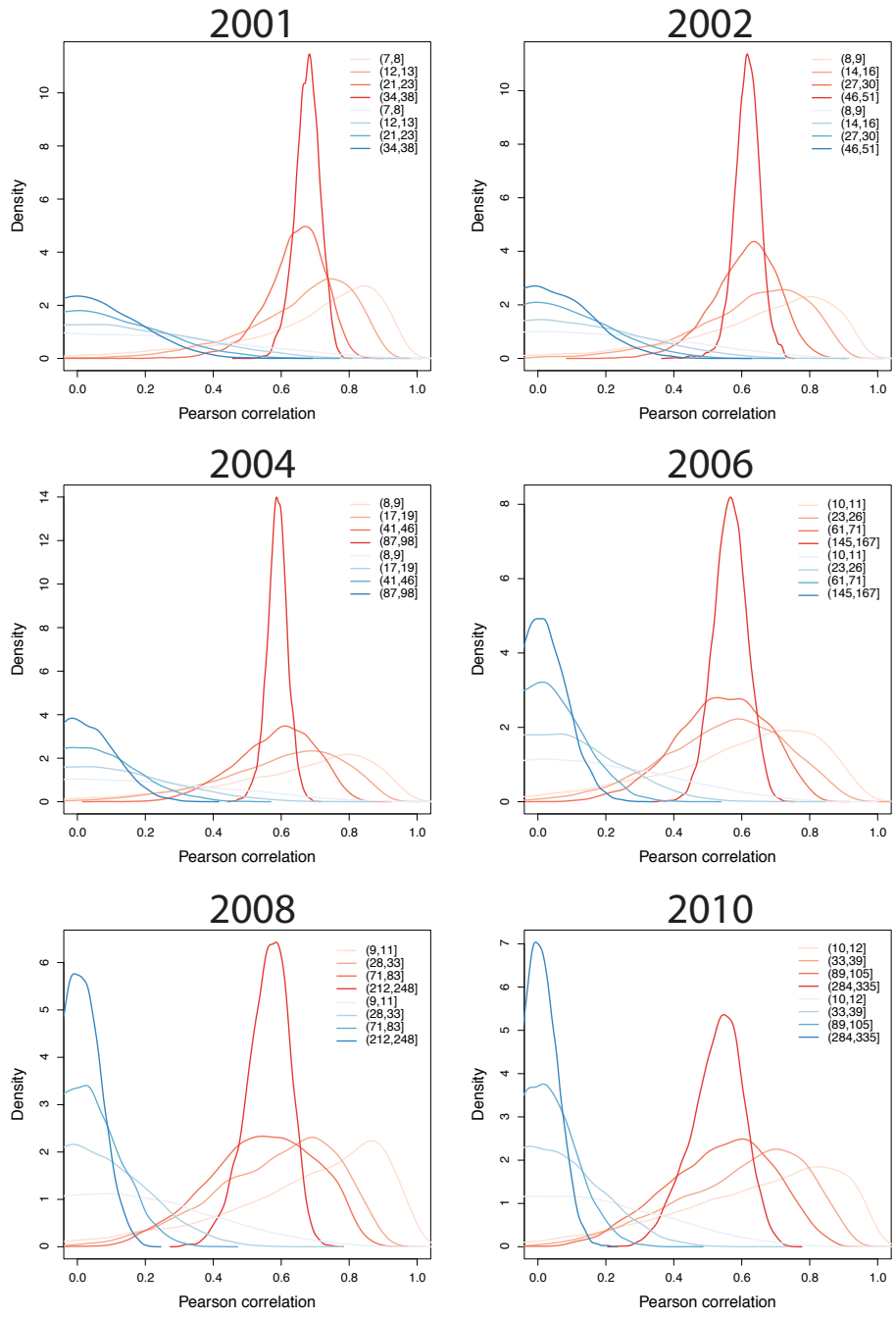


Figure 4.7: **Density functions for the distribution of expected correlation coefficients in sets of random pairs of numbers and sets of distances extracted from pairs of permuted phylogenetic trees**. The genomes available at different time points in the past were used as reference to generate shuffled trees for *E. coli* proteins and the corresponding distributions of tree similarities (red) were calculated for the pairs of trees sharing different numbers of organisms in common (between brackets). Those distributions were compared with equivalent ones generated from random sets of numbers in the same size intervals (blue).

over the years, indicating that it could be independent of the growing redundancy and most likely related with the methodological setup.

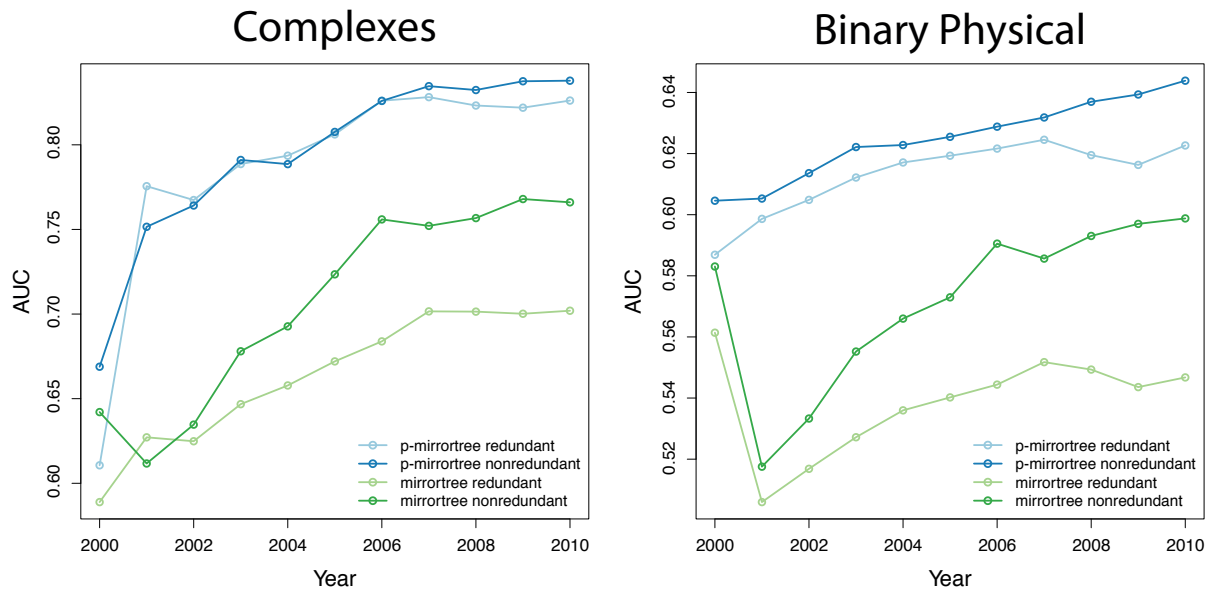


Figure 4.8: **Performance of the MT and pMT methods when predicting physical interactions using different sets of organisms based on the fully-sequenced genomes available in the period 2000–2010 (with and without redundancy)**. The performances were evaluated in terms of AUC using two gold standard datasets of protein interactions: binary physical and co-membership in the same macromolecular complex.

As a consequence of the incomplete *P*-values previously used in MT implementations, some workarounds have been proposed in order to improve global performance. The most common approach is to ignore those protein pairs below a given number of organisms in common. We benchmarked this workaround using MT and pMT to predict proteins in the same macromolecular complex. The same "historical" sets of organisms were evaluated excluding those pairs with less or equal than 15 and 30 organisms (Figure 4.9).

Although MT predictions show better AUCs than those obtained using all the predictions (section 4.3), the performance starts to drop drastically at a certain number of sequenced organisms for both "redundant" and "non-redundant" sets. The maximum performance is different depending on the number of minimum organisms and also with the set of organisms used to construct the phylogenetic trees. For example, using all the sequenced organisms available, the optimal performance of MT was reached in 2003 using tree pairs with more than 15 organisms in common and in 2006 when pairs with more than 30 organisms in common were considered. Disregarding the obvious loss in prediction coverage, as more organisms are used to generate the trees, pairs with larger number of organisms in common need to be excluded in order to get the optimal performance. Even on those maximum values of AUC, MT never outperforms the more stable pMT.

Certain limitations arise as a consequence of all of the above. Firstly, the difficulties to assign the proper threshold of minimum organisms in common necessary to consider a pair of trees constructed with a given set of organisms limits the applicability of MT methodology. As it has

been pointed out, this threshold varies depending on unknown factors related with the size and redundancy of the set of organisms, and thus it remains difficult to define as a single recipe. Secondly, even if the adequate threshold is detected, predicting less protein pairs have a strong negative impact since it can bias the pairs towards proteins conserved in large number of organisms. The drawbacks of excluding protein pairs are not limited to the performance evaluation. Having less pairs evaluated is also critical in those methods, such as PC or CM, that benefit from the whole coevolutionary network to predict single interactions.

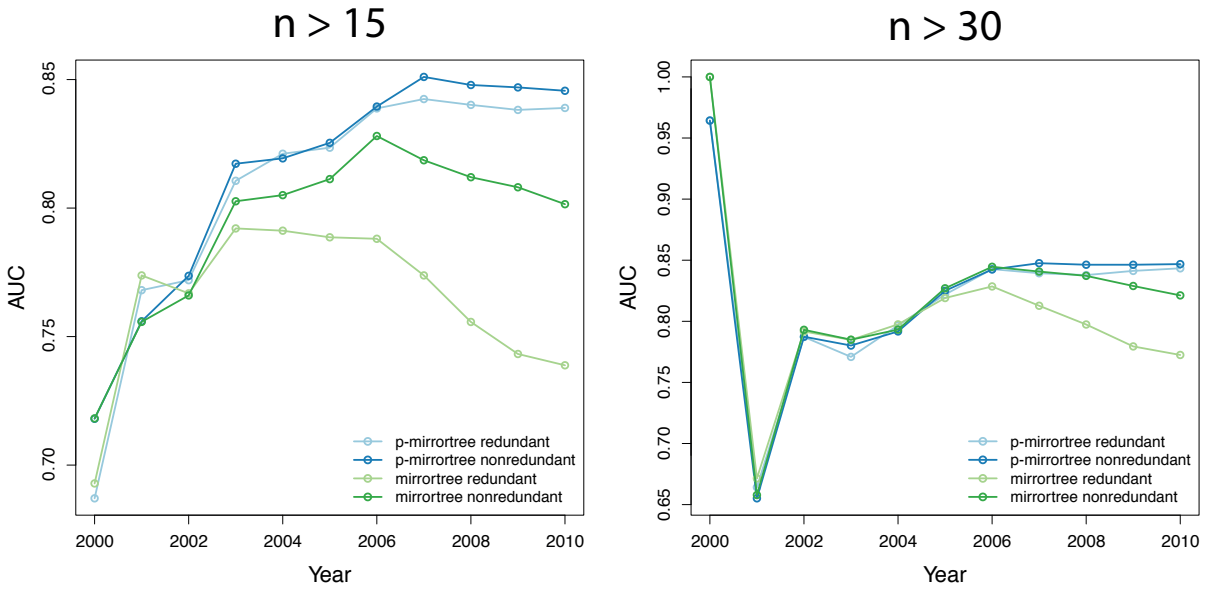


Figure 4.9: **Performance of the MT and pMT methods when predicting proteins in the same macromolecular complex using different sets of organisms based on the fully-sequenced genomes available in the period 2000–2010 (with and without redundancy)**. Their performance was evaluated in terms of AUC. Those pairs of trees with less or equal than 15 or 30 organisms in common were excluded for evaluation.

4.4.3 Context-based *p-mirrortree*

A novel generation of methods succeeded in using the whole matrix of *mirrortree* pairwise correlation coefficients to predict single interactions (PC [164], section 3.1.1.2; and CM, section 3.1.1.3). In order to evaluate the applicability of *p-mirrortree* results to context-based methodologies, we benchmarked PC based on MT correlation coefficients and based on *P*-values. In Figure 4.10, we observe that the PC calculated using *P*-values outperforms the original PC implementation which used correlation coefficients. Considering that the accuracy of pMT are higher than those of MT (section 4.4.2), the improvement in PC predictions can simply arise as a consequence of this.

One of the benefits of these context-based methodologies is the significant reduction in the number of false positives. A new approach, named Hierarchical Co-evolutionary Analysis (HCA), was introduced in order to reassess the similarity between a pair of co-evolutionary profiles. In this approach, a hierarchal clustering was applied over the whole list of co-evolutionary profiles

(section 4.4.3). The purpose of this was to re-score the candidate pairs based on the cophenetic distances in the resulting clustering. By using this score, the enrichment in positive interactions among the first ranked pairs is much larger than using the PC method based on either correlation coefficients or *P*-values (Figure 4.10). Different algorithms for hierarchical clustering showed similar results (Figure S9).

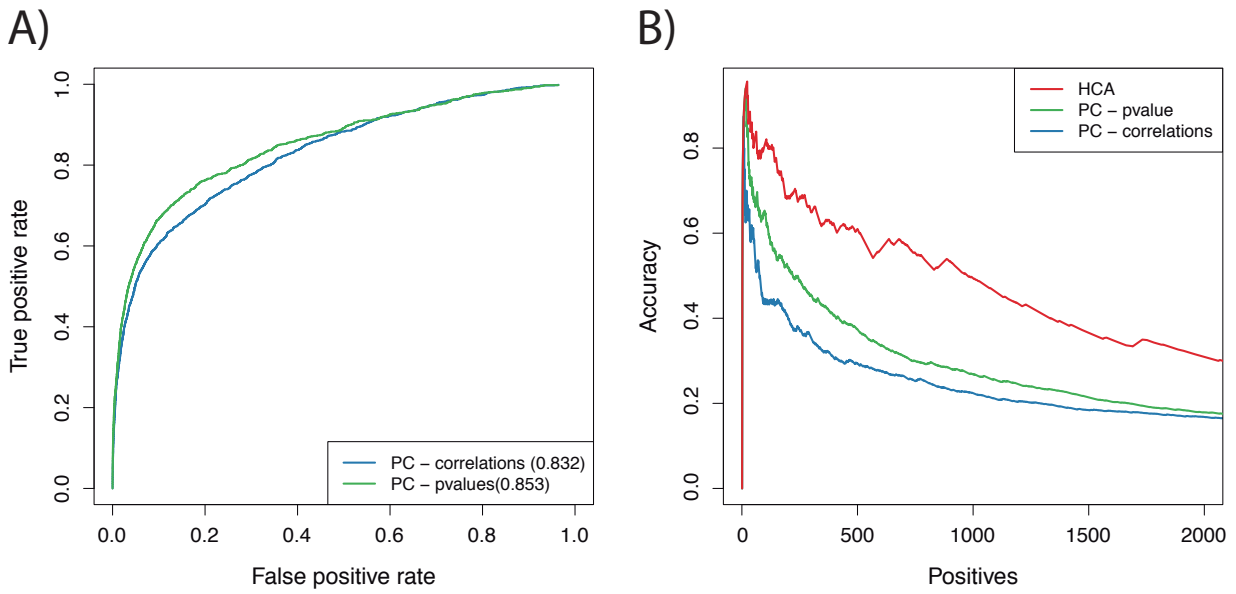


Figure 4.10: **Performances of context-based approaches to predict protein interactions in the "Complexes" gold standard using mirrortree and p-mirrortree scores. A)** ROC curves for PC based on *mirrortree* correlation coefficients - blue - and *p-mirrortree* *P*-values - red. The AUCs of both curves are shown within brackets. **B)** Accuracy *vs.* number of positives in three protein interaction predictions based on PC scores using the matrixes of pairwise correlations - blue - and *P*-values - red -, and cophenetic distances when applying a hierarchal clustering based on Ward's minimum variance over the whole list of co-evolutionary profiles - red.

Apart from its higher performance, an additional advantage to this analysis is that the clustering describes the "co-evolutionary relationships" between a set of proteins in a hierarchical representation, which might be used to infer additional information on the substructure and functioning of the interactome. For example, if we analyze the hierarchical coevolutionary relationships of the *E. coli* ATP synthase (Figure 4.11), we observe how the different members of the complex form clusters, represented as a tree, ranging from the most similar pairs of coevolutionary profiles to the cluster including all the proteins. At each intermediate cut of the tree we can split the proteins into different groups, based on the distances between their coevolutionary profiles. In the case of the ATP synthase, if we cut the tree into three different clusters, we observe a cluster containing the "a" and "c" subunits, a second cluster formed uniquely by the subunit "b" and a third cluster containing the 5 different members of the F_1 particle. These results are compatible with the three-dimensional model of the ATP synthase, in which the "a" and "c" subunits are embedded in the membrane to create the proton pore, the F_1 particle is the cytosolic machinery in charge of the ADP phosphorilation and the subunit "b" connects both sub-complexes (Figure 4.11). Consequently, this coevolutionary analysis generated some clues on the architecture of the macromolecular complex,

only using information at sequence level.

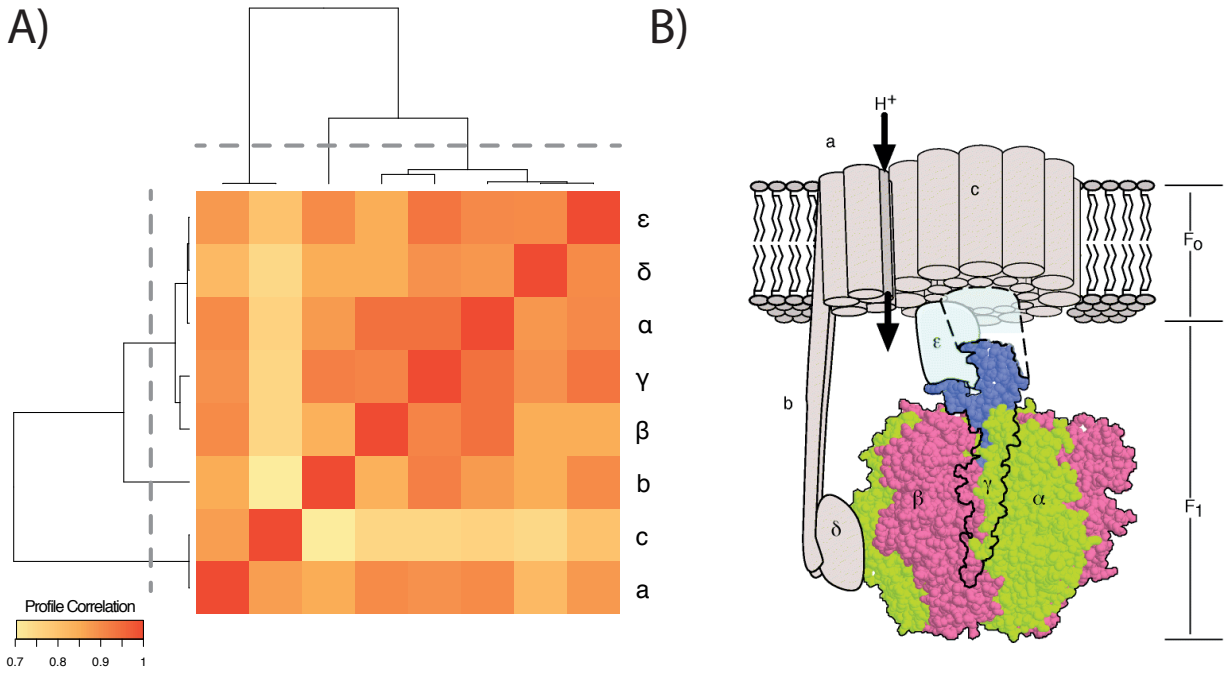


Figure 4.11: **Hierarchical Coevolutionary Analysis (HCA) of the 8 subunits of the *E. coli* ATP synthase**. **A)** Heat map representing the pairwise PC between the coevolutionary profiles of the ATP synthase subunits. The hierarchical clustering is calculated using the Ward's minimum variance algorithm over the pairwise PC results. The clustering is therefore based on the full matrix of pairwise similarities and not only on the similarities shown in the heat map. **B)** Three-dimensional representation of the E.coli ATP synthase based on the model of Wang and Oster [224].

Discussion

In the previous sections, we have reviewed in detail some of the open problems (section 1.7) inherent to the *mirrortree* family of methods. This work tries to diagnose some of these problems and propose alternatives or recipes to overcome some of the methodological limitations that appear as a consequence of our incomplete understanding of the co-evolutionary process affecting interacting proteins. Here, we propose a set of recommendations that can improve the performance and range of applicability of *mirrortree*-based methodologies.

During the last decade, this family of techniques has been largely used by many different groups to predict protein interactions mostly in genome-wide computational experiments. However, the co-evolution-based prediction is also valid when single putative interactions need to be evaluated. Therefore, we present the Mirrortree Server, in which the users with any level of expertise can carefully explore the similarity between a pair of phylogenetic trees using an interactive interface.

More difficulties arise when the co-evolution analysis is performed for large sets of potentially interacting proteins. Some of these problems appear as a consequence of our lack of understanding of the ultimate cause(s) for the observed co-evolution of interacting proteins. Although some authors suggested that co-adaptation and co-evolution occur at the same time in highly dependent proteins, others have shown skeptical regarding the co-adaptative phenomenon. All attempts aimed at getting insight into this by evaluating co-evolution at protein surfaces/interfaces (the regions intuitively related to co-adaptation) used information on crystallized protein complexes, which is scarce. We approach this problem using predicted solvent accessibility, which can be obtained for any sequence with good accuracy.

Other problems arise from the growing number of sequenced organisms available. Whether the large number of redundant genomes affect the *mirrortree*-based methodologies is a capital question necessary to understand the viability of the method in the next years. Therefore, we evaluated the performance of different *mirrortree*-related methods predicting different sets of interactions using trees generated from different sets of organisms selected based on taxonomical criteria.

Finally, considering all the aforementioned analysis, we redesigned the original *mirrortree* to re-score the interacting candidates based on a null distribution of random co-evolution. The distances within a set of phylogenetic trees based on a given set of organisms are randomly permuted to generate a null distribution of tree correlations. This distribution is used to calculate the probability of finding a given correlation coefficient. Using the sets of organisms available in the period 2000–2010, we benchmarked the performance of the original *mirrortree* and the

introduced method, named *p-mirrortree*. Moreover, we studied the ability of this novel metric to predict protein interactions using context-based methodologies, in which the full matrix of pairwise similarities is used for the prediction of single pairs.

In the next sections, we will discuss the results of these analysis and their implications in future improvements of co-evolution-based prediction of protein interactions. Here, we provide different tools and recipes that will enhance the accuracy and understanding of interaction predictions for users with any level of expertise.

5.1 Mirrortree web server

The Mirrortree server represents the first web application for interactively assessing the co-evolution between a pair of phylogenetic trees in a taxonomic framework. Since the tool contains a number of additional functionalities, the server is adequate for either general tree comparisions or co-evolution studies focused on protein interaction prediction.

One of the main advantages of Mirrortree server is its simplicity. The server combines, for the first time, a straight forward pipeline to automatically generate phylogenetic trees with a user-friendly visualizer to explore the results. The multiple dependencies necessary to implement a standalone pipeline made it difficult for non-expert users to perform co-evolution analysis. The server overcomes some tasks, *a priori* difficult for most biologists, such as keeping the databases updated, running programs in command line or parsing files. Any user with a flash featured browser can, starting from two single sequences, compare a pair of protein families of interest without any additional software. But the tool is not only intended for non-expert users. Those users with a deeper knowledge in the field might also tune the phylogenetic reconstruction using the advanced options. Moreover, the Mirortree Server UI provides an intuitive framework to compare a pair of phylogenetic trees and explore their similarities in a taxonomic context. The UI can be used to visualize the trees generated by the server, or any other pair of trees provided by the user, in which the mapping between leaves is known. Because of the interface's flexibility, the applications of this tool, originally designed to explore co-evolution, can be extended to any other areas in which the comparison of phylogenetic trees is informative.

Previous tools have also tried to explore similarities between two or more trees but focusing on different aspects. For example, TSEMA [169] explores the tree similarity between a pair of protein families already reported as interacting. The general objective of this tool is to find the proper mapping between the equivalent protein homologs in both trees. Contrary to the Mirrortree server, this tool requires MSAs as input, hence making it more difficult to use for non-expert users. Moreover, it does not allow to interactively manipulate the trees. Another example of a web server implemented to quantify the similarities in protein evolutionary histories is ADVICE [144]. This tool, intended to postulate potentially interacting pairs, also compares automatically generated phylogenetic trees. However, the results are shown as simple correlation coefficients, ignoring the additional information necessary to understand the results. In that sense, tree representations

and different features to interactively analyze the trees using taxonomical information would add applicability to this tool. Moreover, this server is no longer available.

During the last 36 months in which the server has been online, more than 1.500 unique visitors submitted more than 3.000 jobs from 62 different countries. A successful example was published in the *Nature* journal studying the peroxiredoxin family [143]. By using the server, the authors observed correlated evolution between this family of proteins and the most ancient known clock mechanism: the cyanobacterial Kai proteins. This evidence supports the hypothesis that redox cycles of peroxiredoxins act as circadian clocks in humans. Indeed, they suggest that, contrary to established hypothesis of independent evolution of circadian clocks within different lineages, both systems may share a common ancestor back in the beginning of the aerobic life.

5.2 Global performance of genome-wide predictions

Alternatively to the low throughput analysis of protein-protein interactions, the *mirrortree* family of methods have proven particularly accurate in predicting different types of interactions at a large scale. Although our main goal was to better understand the different phenomena rather than comparing the different variants of the methodology, the results are in agreement with previous studies [164]. The lower MT performance compared with PC and CM had been previously reported but can be better understood at the light of our results. The naive design of MT ignores the phylogenetic biases introduced as consequence of the incomplete statistical model. This problem seems to be corrected by using pMT, which obtains similar AUCs to those algorithms that take advantage of the genome-wide co-evolutionary networks, such as PC, CM or the HCA. Despite of not reaching pMT's AUC, this family of context-based methods is particularly interesting to predict a few number of reliable interactions, since the number of false positives is considerably reduced. More interesting is the ability of HCA to inform about the structure of co-evolving protein complexes.

Another general trend can be observed if we attend to the performance when the predictions are evaluated using different types of interactions. The fact that all methodologies render better results for interactions in the same macromolecular complexes had also been reported [164]. Whether this particularly good performance is a consequence of biological reasons, such as the permanent and stable nature of the interactions between proteins within the same complex, or to technical artifacts, derived from the construction of the negative gold standard, remains unclear. On the other hand, the prediction of binary physical interactions, despite not presenting such high AUCs, presents a reasonable performance, which benefits from the growing number of organisms sequenced over the last decade. The reason for those intermediate AUCs could be related with the heterogeneity of the gold standard, in which the accurately predicted ancient interactions would be mixed with newer interactions presenting similarities in local regions of the trees. Moreover, our results indicate that the predictions of binary physical interactions can be improved by incorporating predicted structural features, such as solvent accessibility. Finally, the prediction of interactions established as co-membership in the same metabolic pathway, in spite of being different from random, might

not be of practical applicability at large scale. The explanation for the observed trends might be that the evolutionary pressure for partners to co-evolve is expected to be higher in proteins forced to interact permanently than in those with occasional associations.

Multiple causes can influence the coverage and AUC of a given prediction such as the number and distribution where the pair of proteins is present, the type of interaction predicted or the scoring method, regardless of the problems related to the results evaluation. Besides, the assessment must be considered in the context of application of a given technique. For example, if we are particularly interested in a small set of highly confident interaction candidates, evaluation using AUC may not be informative; hence, alternative evaluations need to be considered such as precision-recall or evaluation of top scoring results.

5.3 Incorporating information on predicted solvent accessibility

The fact that the incorporation of accessibility predictions worsens MT results suggests that the results of this methodology might be highly influenced by general non-specific co-evolutive processes. These general forces, which can produce similarities in evolutionary rates, would spread their signal through the whole sequences and would not necessarily be restricted to certain regions, such as exposed residues. This observation is supported by the fact that MT performances correlate positively with the number of positions used to generate the trees. In consequence, when the informative residues, spread along the polypeptide chain are removed, the performance of MT predictions decreases.

On the other hand, the PC and CM methods benefit from the use of predicted accessibility when applied to the prediction of binary direct physical interactions. Within the possible explanations for this improvement is that the sensibility of these two methods, previously associated with a more specific co-evolution [210], detects co-adaptative events more frequently than the original *mirrortree*. These co-adaptative events intuitively occur at a higher frequency in binary physical interactions and closer to the interfaces, where the role of compensatory mutations is more critical than in partners without direct physical interaction. Therefore, the enrichment in co-adapting residues expected when filtering by predicted accessibility, together with the higher sensibility of the PC and CM methods to these events, explain the improvement observed in the results.

Another interesting point refers to the definition of accessibility that optimizes the interaction prediction. The results indicate that those residues with a minimal area predicted as accessible (≥ 2Å^2) work better than those with larger areas (≥ 12Å^2 and ≥ 50Å^2). This evidence suggests that the co-adaptative signal is not necessarily restricted to totally exposed residues but can also happen between neighbors through allosteric effect. Thus, in order to optimize the predictions, the adequate number of residues needs to be considered to balance the enrichment of co-adaptative signal *versus* the more general co-evolutionary signals.

The results also indicate that the accessibility predictions obtained from MSAs constructed for this purpose yield better performances than those predicted from the MSAs used to generate

the phylogenetic trees. The fact that "richer" alignments containing eukaryotic sequences and paralogs show more accurate predicted accessibilities was corroborated by previous reports in which the authors state that the quality of the MSA is critical for obtaining good accessibility predictions [225]. They suggest that accessibility predictions are more sensitive to alignment errors than other features such as secondary structure. This observation might be consequence of the fact that accessibility is evolutionarily less conserved than the secondary structure [226]. Unfortunately, the accessibilities predicted from the already generated MSAs of orthologs show sub-optimal performances, and thus different MSAs need to be generated *ad hoc* for this purpose.

5.4 Co-evolution *vs* Co-adaptation

The observation that predicted accessibility only helps the PC and CM predictions in detecting binary physical interactions can be also framed in the co-adaptation *vs.* general co-evolution debate.

As commented before, the causes underlying the observed co-evolution between interacting proteins are still not totally clear. There is some controversy in the field to define what is observable at sequence level versus the underlying evolutionary phenomenon. Here we remain tight to the disambiguation between the clearly observable phenomenon of "co-evolution" and the more elusive concept of "co-adaptation" between co-evolving components [141, 210, 227]. In this definition, co-evolution would be confined to the general concerted patterns of co-variation observed at sequence level without implying reciprocal evolutionary events. Alternatively, co-adaptation implies reciprocity as the causal phenomenon behind the observed co-evolution. In this sense, co-adaptation would be referring to, for example, the compensatory changes required for maintaining protein interactions. The more general co-evolution, which usually implies a similarity in the evolutionary rates, has been described for proteins which do not necessarily interact physically, but share a certain functional relationship [180, 187]. Although these general forces seems to dominate the tree-tree similarities, compensatory changes have been repeatedly observed in protein interfaces and are surely playing a role in the co-evolution of interacting proteins. However, it is difficult to conceive that these local compensatory mutations are enough to substantially affect the distances in a pair of phylogenetic trees and thus their global similarity. Previous studies aimed at getting insight into this phenomenon focusing on interfaces or the surrounding surfaces of structurally determined interactions, in order to enrich the signal in regions amenable of compensatory mutations [176, 177]. Nevertheless, both studies were limited by the relatively small amount of structural data available, so the true extent of their affirmations needs to be evaluated.

At the light of the results described herein, we demonstrated that protein interaction prediction can be improved by using predicted solvent accessibility under certain circumstances. Unsurprisingly, those methods detecting more specific similarities - PC and CM - were the ones achieving greater performances when predicting those type of interactions in which the co-adaptation is expected to play a more important role: the binary physical interactions. These results suggest

that a co-adaptative signal, despite being residual, can be informative in order to predict physical interactions under certain conditions.

5.5 Selection of organisms

Our results unambiguously indicate that the set of organisms used to generate the phylogenetic trees critically influences the protein interaction predictions. In general, the growing number of available organisms produces two opposite consequences in *mirrortree*-based predictions. On one hand, the phylogenetic reconstructions are then based on more complete representations of the evolutionary history of the protein families. Therefore, it is not surprising that the methods benefit from the enriched co-evolutive information available when comparing more detailed trees. On the other hand, the sequencing efforts have explored the tree of life in a non-homogenous way. Whereas several redundant organisms have been sequenced in some clades, others remain relatively incomplete, artificially biasing the co-evolutionary analysis. As a consequence, we observe that MT predictions perform better when using complete and non-redundant representations of the tree of life. Indeed, as long as more organisms are sequenced, the accumulation of redundant organisms increases the performance gap between the predictions calculated using all the available organisms and the ones obtained using an equivalent list where the redundant organisms have been excluded. Alternatively, PC, CM and pMT deal better with this redundancy bias, showing similar performances in equivalent redundant and non-redundant sets.

Once we understand that the set of organisms used to generate the phylogenetic trees influences the interaction prediction, it is necessary to pay attention on how the taxonomical distribution of these organisms influences the prediction of the different types of interactions. Ancient and stable physical interactions might be more accurately predicted using sets of organisms different than the ones used to predict transient or functional interactions. Ancient interactions such as those happening between proteins in the same macromolecular complex, for instance, are expected to be conserved for most of the orthologs along the tree of life, hence the associated co-evolutive landmark is expected to be spread through the whole taxonomy. Indeed, this hypothesis is supported by our results, since we observe better performances when proteins in the same complex are predicted using organisms spread along the whole taxonomy. The opposite phenomenon is observed when we evaluate binary physical interactions. These interactions are better predicted when we use subsets of taxa close to the reference organism. In general, binary physical interactions are "newer" and enriched in transient interactions. Considering that "rewiring" is more frequent within transient interactions [228], it might occur that the orthologs of two proteins involved in a transient interaction in a given organisms are not interacting in a relatively distant one [201, 229]. In other words, many of the interactions we are evaluating might be new and hence specific for *E. coli* and its close neighbors. This hypothesis would explain that our predictions for these type of interactions are better when these particular genomes are used for constructing the trees. Similar improvement is shown when the analysis of proteins involved in the same metabolic pathway is limited to close taxonomical neighbors. Interestingly, similar relationships between the "age"

of the interactions, their conservation across the taxonomy and the optimal set of organisms for predicting them has been reported for the "phylogenetic profiling" method [230].

The comparison of distance matrixes is a NP-hard problem and, consequently, can be severely affected as the number of available genomes continues to grow. Even robust methods such as pMT need to solve the computational problem of calculating millions of branch lengths for every phylogenetic tree. Hence, our results invite to propose a set of simple and general "recipes" on which sets of organisms more properly detect the co-evolutionary forces behind a given type of interaction, avoiding to use all available. The first question that arises when a user is about to perform a regular co-evolution-based prediction of protein interactions is if a set of phylogenetic trees, which can be considered as background, is available (i.e. all the proteins in the same organisms, proteins in the same cellular compartment). With a set of background phylogenetic trees, the statistical confidence of a given correlation can be evaluated (pMT), or the coevolutionary profiles considered as proxies of the co-evolutionary signal of a given protein (PC and CM). All these methodologies are particularly robust to the set of organisms and benefit from rich representations of the evolutionary histories of the proteins. Alternatively, in order to reduce the computational cost of the analysis, the user can restrict the study to the non-redundant organisms without expecting a negative impact in the prediction performance. On the other hand, the requisites to obtain reasonable performances in MT are more restrictive. The set of organisms used to generate the phylogenetic trees should be filtered by taxonomic redundancy. Filtering at the strain or species level seems to be enough according to our results.

5.6 Evolutionary significant assessment of co-evolution

The results presented here demonstrate that the prediction of protein interactions is clearly improved when the statistical confidence of the correlation is evaluated based on a background distribution of tree similarities. Using this distribution of expected correlations corrects in a natural way many of the factors discussed that affect the performance of the original MT, including the background similarity derived of the underlying speciation process [132, 160], the redundancy of the original set of organisms and the different range of organisms in which the candidate proteins are present.

One of the main observations previously described is the fact that the set of organisms used to generate the phylogenetic trees conditions the resulting correlation. Indeed, one of the main advantages of pMT compared with methods such as PC or CM is its ability to evaluate independently pairs of proteins sharing different number of organisms in common. Since the analysis of all the possible combinations of organisms in common would be unfeasible due to its computational complexity, pMT uses the number of organisms in common as an approximation. As the phylogenetic trees of a given protein are usually generated looking for orthologs using an organism as a reference, the number of organisms shared by a couple of trees is directly related to the set of organisms in which both proteins are present. Despite the genes sharing events of horizontal gene transfer, it is expected to find the different sized groups enriched in the same organisms.

Another advantage of pMT over MT predictions relies on its coverage. Usually, MT is applied only to the pairs of trees with a minimum number of organisms in common and, consequently, many pairs of families are not explored. This common procedure significantly reduces the method's coverage diminishing the final results. Since the pMT *P*-values can be calculated for any pair of trees regardless the number of organisms in common, the improvement in AUC is produced at no cost in terms of coverage. Indeed, the full coverage of pMT is particularly interesting for context-based approaches such as PC or CM, which take as input the whole matrix of co-evolutionary scores. Alternatively to the sparse matrixes produced by MT, pMT results contain the full matrix of possible pairs in the genome of interest. Considering the better AUC and the better coverage, it is not a surprise that PC obtains better accuracies using pMT *P*-values than MT correlations. More promising are the results proposed by the here introduced HCA methodology, which considerably reduces the number of false positives. Moreover, the HCA can be informative of some structural features of the protein complexes, such as the example of the ATP synthase. This information is extremely valuable considering that the method is able to describe that kind of structural relationships only using information at sequence level.

5.7 Future Developments and Perspectives

The promising results described here leave the door open to future improvements on co-evolution-based prediction of protein interactions. Whereas some research is necessary to polish the technical details on creating and comparing accurate phylogenetic trees, a deeper effort is needed in order to capture the co-dependencies at the sequence level established between interacting proteins.

Although the tree reconstruction has not been shown to be a limiting step on protein interaction prediction, a proper tuning of the technical details might help the detection of co-evolutionary events. The intermediate steps necessary to accomplish this task, such as ortholog detection, sequence alignment, distance estimation or tree generation, need to be carefully considered, particularly in complex evolutionary scenarios such as HGT events [132]. Future implementations with a reasonable tradeoff of computing time and accuracy might also help to build more reliable evolutionary models at a large scale.

Another unresolved problem is the measurement of the phylogenetic tree similarity. As the tree comparison requires unambiguous mapping of sequences at a species level, the "correct" protein orthologs need to be identified. This task can be especially hard in eukaryotes in which the large number of paralogs and multi-domain proteins complicate the correct assignment [165]. On the other hand, the correlation of protein distances calculated either from the MSAs [130] or from the phylogenetic trees [132] has proven particularly useful as a similarity metric, specially in the framework of the newer methods developed here. Nevertheless, some studies suggest that most complex schemas involving multidimensional spaces are needed in order to consider the multiple dependencies of protein distances [162].

Another limitation to the progress in this field is the need to correctly benchmark the results of the predictive methods. The lack of adequate gold standards for positive and negative interactions

conditions the evaluation of the methods (section 1.6.6) especially for large-scale predictions. In the future, blind tests with hidden gold standards (similar to the CASP contests in protein structure prediction) would be the best way to accurately compare the different methods.

One of the emerging problems, which will gain importance on the next years, lies on the consequences derived of the huge amount of genomic data generated nowadays. On one hand, we know with this work that, with the proper methods, the interaction prediction still improves as more and more genomes become available. On the other, the computational complexity growths exponentially with the number of genomes requiring, large computational resources. In the future, heuristic methods similar to the ones looking for optimal sets of organisms in phylogenetic profiling [118, 119] might help to reduce the complexity and avoid a waste of computational time in phylogenetic tree comparison.

This work has also practical implications for the application of these methodologies, and these are not only related to the general improvement in the prediction of protein interaction. It opens interesting possibilities for studying how the exposed residues change and co-adapt during evolution. This could give some insight into the physico-chemical basis of protein interactions, since the coordinated changes at these regions would provide a picture of possible changes affecting the interactions. These residues might be good candidates for mutagenesis experiments aimed at switching the interaction specificity of the proteins and/or adapting them to new interaction partners.

Moreover, the Mirrortree Server framework should serve as reference for highly specialized groups studying in parallel the co-evolution of their respective systems. Hence, it is expected that our understanding of the co-evolution at a molecular level increases with multiple examples such as the peroxiredoxins [143].

Conclusions

1. We have developed a system, MirrorTree Server, to study the co-evolution of protein families in a taxonomic context. This tool includes an automatic pipeline to generate the phylogenetic trees of the protein families of interest, and an interactive visualizer to explore their similarities and taxonomic characteristics. Due to the multiple functionalities it incorporates, the server has proven to be useful for users with any level of expertise.

2. Globally, our results reinforce previous observations in which the *mirrortree*-based family of methods proved being particularly accurate in predicting different types of interactions at a large scale. Moreover, important information on the structure, function, and evolution of interacting proteins can be inferred using these methodologies.

3. Incorporation of predicted solvent accessibility helps the co-evolution based prediction of binary physical interactions when context-based methodologies such as *Profile Correlation* or *ContextMirror* are applied.

4. The set of organisms used to generate the phylogenetic trees conditions the final performance of the protein interaction prediction methods based on co-evolution. In general, the inclusion of sequences belonging to close organisms helps the prediction of "recent" interactions, whereas sequences belonging to distant organisms help the prediction of relatively "ancient" interactions.

5. We introduced a new methodology denominated *p-mirrortree* to evaluate the protein co-evolution between a pair of phylogenetic trees based on a null distribution of tree similarities obtained by shuffling a background set of phylogenetic trees. This methodology outperforms the original *mirrortree*, independently of the set of organisms used to generate the phylogenetic trees. Moreover, *p-mirrortree* results demonstrated a higher accuracy when used as input of context-based approaches.

Conclusiones

1. Hemos desarrollado un servidor, MirrorTree Server, para el estudio de la coevolución de familias de proteínas en un contexto taxonómico. Esta herramienta incluye un flujo de trabajo para la generación automática de árboles filogenéticos de las familias de proteínas de interés y un visualizador para explorar sus similitudes y características taxonómicas. Debido a las múltiples funcionalidades incorporadas, el servidor ha demostrado ser útil para usuarios con distintos niveles de experiencia en el área.

2. Nuestros resultados refuerzan globalmente las observaciones previas en las cuales la familia de métodos basada en *mirrortree* se muestra especialmente precisa en la predicción de de distintos tipos de interacciones a gran escala. Además, información relevante en cuanto a la estructura, función y evolución de las proteínas puede ser extraída de los resultados de estas metodologías.

3. La incorporación de datos de accesibilidad predichos ayudan a la predicción de interacciones binarias basadas en coevolución cuando se emplean métodos basados en contexto como *Profile Correlation* o *ContextMirror*.

4. El conjunto de organismos usado para la generación de los árboles filogenéticos condiciona la capacidad de los métodos para predecir interacciones basadas en coevolución. En general, la inclusión de secuencias pertenecientes a organismos cercanos filogenéticamente puede ayudar en la detección de interacciones “recientes”, mientras que la inclusión de secuencias pertenecientes a organismos lejanos ayuda a la predicción de interacciones “antiguas”.

5. Presentamos una nueva metodología denominada *p-mirrortree* capaz de evaluar la coevolución de un par de árboles filogenéticos basándose en una distribución nula de similitudes de árboles obtenida como consecuencia de barajar un conjunto de árboles de referencia. Esta metodología mejora el *mirrortree* original independientemente del conjunto de organismos empleados para generar los árboles filogenéticos. Además, los resultados de *p-mirrortree* han demostrado una mayor precisión cuando son usados como entrada para los métodos basados en contexto.

Bibliography

[1] Hiroaki Kitano. Systems biology: a brief overview. *Science (New York, NY)*, 295(5560): 1662–1664, March 2002.

[2] Paul Nurse. Systems biology: understanding cells. *Nature*, 424(6951):883, August 2003.

[3] S H Strogatz. Exploring complex networks. *Nature*, 410(6825):268–276, March 2001.

[4] S Wasserman and K Faust. Social Network Analysis: Methods and Applications - Stanley Wasserman, Katherine Faust - Google Books. 1994.

[5] J Scott. Social Network Analysis: A Handbook - John Scott. 2000.

[6] D J Watts and S H Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393 (6684):440–442, June 1998.

[7] R J Williams and N D Martinez. Simple rules yield complex food webs. *Nature*, 404(6774): 180–183, March 2000.

[8] Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. Graph structure in the Web. *Computer Networks*, 33(1-6):309–320, June 2000.

[9] Albert-Laszlo Barabasi and R Albert. Emergence of scaling in random networks. *Science (New York, NY)*, 286(5439):509–512, 1999.

[10] Albert-Laszlo Barabasi and Zoltan N Oltvai. Network biology: understanding the cell's functional organization. *Nat Rev Genet*, 5(2):101–113, 2004.

[11] S Milgram. The small world problem. *Psychology today*, 1967.

[12] H Jeong, B Tombor, R Albert, Z N Oltvai, and Albert-Laszlo Barabasi. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, 2000.

[13] R Milo, S Shen-Orr, S Itzkovitz, N Kashtan, D Chklovskii, and Uri Alon. Network motifs: simple building blocks of complex networks. *Science (New York, NY)*, 298(5594):824–827, October 2002.

[14] E Golemis and P D Adams. *Protein-Protein Interactions: A Molecular Cloning Manual*. Cold Spring Harbor Laboratory Press, 2005.

[15] H M Berman, T N Bhat, P E Bourne, Z Feng, G Gilliland, H Weissig, and J Westbrook. The Protein Data Bank and the challenge of structural genomics. *Nature structural biology*, 7 Suppl:957–959, November 2000.

[16] Jennifer Lippincott-Schwartz and George H Patterson. Development and use of fluorescent protein markers in living cells. *Science (New York, NY)*, 300(5616):87–91, April 2003.

[17] Jacob Piehler. New methodologies for measuring protein interactions in vivo and in vitro. *Curr Opin Struct Biol*, 15(1):4–14, February 2005.

[18] Koon-Kiu Yan, Gang Fang, Nitin Bhardwaj, Roger P Alexander, and Mark Gerstein. Comparing genomes to computer operating systems in terms of the topology and evolution of their regulatory control networks. *Proceedings of the National Academy of Sciences of the United States of America*, 107(20):9186–9191, May 2010.

[19] Robert Karlsson. SPR for molecular interaction analysis: a review of emerging application areas. *Journal of molecular recognition : JMR*, 17(3):151–161, May 2004.

[20] Matthew A Cooper. Label-free screening of bio-molecular interactions. *Analytical and bioanalytical chemistry*, 377(5):834–842, November 2003.

[21] Adrián Velázquez Campoy and Ernesto Freire. ITC in the post-genomic era...? Priceless. *Biophysical chemistry*, 115(2-3):115–124, April 2005.

[22] Yong Yang, Hong Wang, and Dorothy A Erie. Quantitative characterization of biomolecular assemblies and interactions using atomic force microscopy. *Methods (San Diego, Calif.)*, 29 (2):175–187, February 2003.

[23] M Margittai, J Widengren, E Schweinberger, G F Schröder, S Felekyan, E Haustein, M König, D Fasshauer, H Grubmüller, R Jahn, and C A M Seidel. Single-molecule fluorescence resonance energy transfer reveals a dynamic equilibrium between closed and open conformations of syntaxin 1. *Proceedings of the National Academy of Sciences of the United States of America*, 100(26):15516–15521, December 2003.

[24] S Fields and O Song. A novel genetic system to detect protein-protein interactions. *Nature*, 340(6230):245–246, July 1989.

[25] S J SJ Fashena, I I Serebriiskii, and E A EA Golemis. The continued evolution of two-hybrid screening approaches in yeast: how to outwit different preys with different baits. *Gene*, 250 (1-2):1–14, May 2000.

[26] Barry Causier. Studying the interactome with the yeast two-hybrid system and mass spectrometry. *Mass spectrometry reviews*, 23(5):350–367, September 2004.

[27] Daniel Auerbach, Safia Thaminy, Michael O Hottiger, and Igor Stagljar. The post-genomic era of interactive proteomics: facts and perspectives. *Proteomics*, 2(6):611–623, June 2002.

[28] Wim Van Criekinge and Rudi Beyaert. Yeast Two-Hybrid: State of the Art. *Biological procedures online*, 2:1–38, October 1999.

[29] G G Toby and E A Golemis. Using the yeast interaction trap and other two-hybrid-based approaches to study protein-protein interactions. *Methods (San Diego, Calif.)*, 24(3):201–217, July 2001.

[30] Jae Woon Lee and Soo-Kyung Lee. Mammalian two-hybrid assay for detecting protein-protein interactions in vivo. *Methods in molecular biology (Clifton, NJ)*, 261:327–336, 2004.

[31] A J Walhout and M Vidal. High-throughput yeast two-hybrid assays for large-scale protein interaction mapping. *Methods (San Diego, Calif.)*, 24(3):297–306, July 2001.

[32] A Aronheim, E Zandi, H Hennemann, S J Elledge, and M Karin. Isolation of an AP-1 repressor by a novel method for detecting protein-protein interactions. *Molecular and cellular biology*, 17(6):3094–3102, June 1997.

[33] W A Mohler and H M Blau. Gene expression and cell fusion analyzed by lacZ complementation in mammalian cells. *Proceedings of the National Academy of Sciences of the United States of America*, 93(22):12423–12427, October 1996.

[34] P L Bartel, J A Roecklein, D SenGupta, and S Fields. A protein linkage map of Escherichia coli bacteriophage T7. *Nature genetics*, 12(1):72–77, January 1996.

[35] R L Finley and R Brent. Interaction mating reveals binary and ternary connections between Drosophila cell cycle regulators. *Proceedings of the National Academy of Sciences of the United States of America*, 91(26):12980–12984, December 1994.

[36] A J Walhout, R Sordella, X Lu, J L Hartley, G F Temple, M A Brasch, N Thierry-Mieg, and M Vidal. Protein interaction mapping in C. elegans using proteins involved in vulval development. *Science (New York, NY)*, 287(5450):116–122, January 2000.

[37] G Rigaut, A Shevchenko, B Rutz, M Wilm, M Mann, and B Séraphin. A generic protein purification method for protein complex characterization and proteome exploration. *Nature biotechnology*, 17(10):1030–1032, October 1999.

[38] O Puig, F Caspary, G Rigaut, B Rutz, E Bouveret, E Bragado-Nilsson, M Wilm, and B Séraphin. The tandem affinity purification (TAP) method: a general procedure of protein complex purification. *Methods (San Diego, Calif.)*, 24(3):218–229, July 2001.

[39] Alessandra Di Tullio, Samantha Reale, and Francesco De Angelis. Molecular recognition by mass spectrometry. *Journal of mass spectrometry : JMS*, 40(7):845–865, July 2005.

[40] Ruedi Aebersold and Matthias Mann. Mass spectrometry-based proteomics. *Nature*, 422 (6928):198–207, March 2003.

[41] C M Whitehouse, R N Dreyer, M Yamashita, and J B Fenn. Electrospray interface for liquid chromatographs and mass spectrometers. *Analytical chemistry*, 57(3):675–679, March 1985.

[42] U Pieles, W Zürcher, M Schär, and H E Moser. Matrix-assisted laser desorption ionization time-of-flight mass spectrometry: a powerful tool for the mass and sequence analysis of natural and modified oligonucleotides. *Nucleic Acids Res*, 21(14):3191–3196, July 1993.

[43] M Karas and F Hillenkamp. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Analytical chemistry*, 60(20):2299–2301, October 1988.

[44] J R Yates, J K Eng, A L McCormack, and D Schieltz. Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Analytical chemistry*, 67(8):1426–1436, April 1995.

[45] J A Taylor and R S Johnson. Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. *Rapid communications in mass spectrometry : RCM*, 11(9): 1067–1075, 1997.

[46] P A Pevzner, V Dancík, and C L Tang. Mutation-tolerant protein identification by mass spectrometry. *Journal of computational biology : a journal of computational molecular cell biology*, 7(6):777–787, 2000.

[47] Lewis Y Geer, Sanford P Markey, Jeffrey A Kowalak, Lukas Wagner, Ming Xu, Dawn M Maynard, Xiaoyu Yang, Wenyao Shi, and Stephen H Bryant. Open mass spectrometry search algorithm. *Journal of proteome research*, 3(5):958–964, September 2004.

[48] S L Rutherford. From genotype to phenotype: buffering mechanisms and the storage of genetic information. *BioEssays : news and reviews in molecular, cellular and developmental biology*, 22(12):1095–1105, December 2000.

[49] J L Hartman, B Garvik, and L Hartwell. Principles for the buffering of genetic variation. *Science (New York, NY)*, 291(5506):1001–1004, February 2001.

[50] A Bender and J R Pringle. Use of a screen for synthetic lethal and multicopy suppressee mutants to identify two new genes involved in morphogenesis in Saccharomyces cerevisiae. *Molecular and cellular biology*, 11(3):1295–1305, March 1991.

[51] Siew Loon Ooi, Xuewen Pan, Brian D Peyser, Ping Ye, Pamela B Meluh, Daniel S Yuan, Rafael A Irizarry, Joel S Bader, Forrest A Spencer, and Jef D Boeke. Global synthetic-lethality analysis and yeast functional profiling. *Trends Genet*, 22(1):56–63, January 2006.

[52] James A Brown, Gavin Sherlock, Chad L Myers, Nicola M Burrows, Changchun Deng, H Irene Wu, Kelly E McCann, Olga G Troyanskaya, and J Martin Brown. Global analysis of gene function in yeast by quantitative phenotypic profiling. *Mol Syst Biol*, 2:2006.0001, 2006.

[53] G MacBeath and S L Schreiber. Printing proteins as microarrays for high-throughput function determination. *Science (New York, NY)*, 289(5485):1760–1763, September 2000.

[54] H Zhu, M Bilgin, R Bangham, D Hall, A Casamayor, P Bertone, N Lan, R Jansen, S Bidlingmaier, T Houfek, T Mitchell, P Miller, R A Dean, M Gerstein, and M Snyder. Global analysis of protein activities using proteome chips. *Science (New York, NY)*, 293(5537):2101–2105, September 2001.

[55] Richard B Jones, Andrew Gordus, Jordan A Krall, and Gavin MacBeath. A quantitative protein interaction network for the ErbB receptors using protein microarrays. *Nature*, 439 (7073):168–174, January 2006.

[56] G Smith. Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science (New York, NY)*, 228(4705):1315–1317, June 1985.

[57] Amy Hin Yan Tong, Becky Drees, Giuliano Nardelli, Gary D Bader, Barbara Brannetti, Luisa Castagnoli, Marie Evangelista, Silvia Ferracuti, Bryce Nelson, Serena Paoluzi, Michele Quondam, Adriana Zucconi, Christopher W V Hogue, Stanley Fields, Charles Boone, and Gianni Cesareni. A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science (New York, NY)*, 295(5553):321–324, January 2002.

[58] J C Rain, L Selig, H de Reuse, V Battaglia, C Reverdy, S Simon, G Lenzen, F Petel, J Wojcik, V Schachter, Y Chemama, A Labigne, and P Legrain. The protein-protein interaction map of Helicobacter pylori. *Nature*, 409(6817):211–215, January 2001.

[59] Gareth Butland, José Manuel Peregrín-Alvarez, Joyce Li, Wehong Yang, Xiaochun Yang, Veronica Canadien, Andrei Starostine, Dawn Richards, Bryan Beattie, Nevan Krogan, Michael Davey, John Parkinson, Jack Greenblatt, and Andrew Emili. Interaction network containing conserved and essential protein complexes in Escherichia coli. *Nature*, 433 (7025):531–537, February 2005.

[60] T Ito, K Tashiro, S Muta, R Ozawa, T Chiba, M Nishizawa, K Yamamoto, S Kuhara, and Y Sakaki. 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*, 97(3):1143–1147, 2000.

[61] P Uetz, L Giot, G Cagney, T A Mansfield, R S Judson, J R Knight, D Lockshon, V Narayan, M Srinivasan, P Pochart, A Qureshi-Emili, Y Li, B Godwin, D Conover, T Kalbfleisch, G Vijayadamodar, M Yang, M Johnston, S Fields, and J M Rothberg. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, 403(6770):623–627, 2000.

[62] B Schwikowski, P Uetz, and S Fields. A network of protein-protein interactions in yeast. *Nature biotechnology*, 18(12):1257–1261, December 2000.

[63] Anne-Claude Gavin, Markus Bösche, Roland Krause, Paola Grandi, Martina Marzioch, Andreas Bauer, Jörg Schultz, Jens M Rick, Anne-Marie Michon, Cristina-Maria Cruciat, Marita Remor, Christian Höfert, Malgorzata Schelder, Miro Brajenovic, Heinz Ruffner, Alejandro Merino, Karin Klein, Manuela Hudak, David Dickson, Tatjana Rudi, Volker Gnau, Angela Bauch, Sonja Bastuck, Bettina Huhse, Christina Leutwein, Marie-Anne Heurtier, Richard R Copley, Angela Edelmann, Erich Querfurth, Vladimir Rybin, Gerard Drewes, Manfred Raida, Tewis Bouwmeester, Peer Bork, Bertrand Seraphin, Bernhard Kuster, Gitte Neubauer, and Giulio Superti-Furga. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868):141–147, January 2002.

[64] Yuen Ho, Albrecht Gruhler, Adrian Heilbut, Gary D Bader, Lynda Moore, Sally-Lin Adams, Anna Millar, Paul Taylor, Keiryn Bennett, Kelly Boutilier, Lingyun Yang, Cheryl Wolting, Ian Donaldson, Søren Schandorff, Juanita Shewnarane, Mai Vo, Joanne Taggart, Marilyn Goudreault, Brenda Muskat, Cris Alfarano, Danielle Dewar, Zhen Lin, Katerina Michalickova, Andrew R Willems, Holly Sassi, Peter A Nielsen, Karina J Rasmussen, Jens R Andersen, Lene E Johansen, Lykke H Hansen, Hans Jespersen, Alexandre Podtelejnikov, Eva Nielsen, Janne Crawford, Vibeke Poulsen, Birgitte D Sørensen, Jesper Matthiesen, Ronald C Hendrickson, Frank Gleeson, Tony Pawson, Michael F Moran, Daniel Durocher, Matthias Mann, Christopher W V Hogue, Daniel Figeys, and Mike Tyers. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature*, 415(6868):180–183, January 2002.

[65] Nicolas Simonis, Jean-François Rual, Anne-Ruxandra Carvunis, Murat Tasan, Irma Lemmens, Tomoko Hirozane-Kishikawa, Tong Hao, Julie M Sahalie, Kavitha Venkatesan, Fana Gebreab, Sebiha Cevik, Niels Klitgord, Changyu Fan, Pascal Braun, Ning Li, Nono Ayivi-Guedehoussou, Elizabeth Dann, Nicolas Bertin, David Szeto, Amélie Dricot, Muhammed A Yildirim, Chenwei Lin, Anne-Sophie de Smet, Huey-Ling Kao, Christophe Simon, Alex Smolyar, Jin Sook Ahn, Muneesh Tewari, Mike Boxem, Stuart Milstein, Haiyuan Yu, Matija Dreze, Jean Vandenhaute, Kristin C Gunsalus, Michael E Cusick, David E Hill, Jan Tavernier, Frederick P Roth, and Marc Vidal. Empirically controlled mapping of the Caenorhabditis elegans protein-protein interactome network. *Nature methods*, 6(1):47–54, January 2009.

[66] L Giot, J S Bader, C Brouwer, A Chaudhuri, B Kuang, Y Li, Y L Hao, C E Ooi, B Godwin, E Vitols, G Vijayadamodar, P Pochart, H Machineni, M Welsh, Y Kong, B Zerhusen, R Malcolm, Z Varrone, A Collis, M Minto, S Burgess, L McDaniel, E Stimpson, F Spriggs, J Williams, K Neurath, N Ioime, M Agee, E Voss, K Furtak, R Renzulli, N Aanensen, S Carrolla, E Bickelhaupt, Y Lazovatsky, A DaSilva, J Zhong, C A Stanyon, R L Finley, K P White, M Braverman, T Jarvie, S Gold, M Leach, J Knight, R A Shimkets, M P McKenna, J Chant, and J M Rothberg.

A protein interaction map of Drosophila melanogaster. *Science (New York, NY)*, 302(5651): 1727–1736, December 2003.

[67] Jean-François Rual, Kavitha Venkatesan, Tong Hao, Tomoko Hirozane-Kishikawa, Amélie Dricot, Ning Li, Gabriel F Berriz, Francis D Gibbons, Matija Dreze, Nono Ayivi-Guedehoussou, Niels Klitgord, Christophe Simon, Mike Boxem, Stuart Milstein, Jennifer Rosenberg, Debra S Goldberg, Lan V Zhang, Sharyl L Wong, Giovanni Franklin, Siming Li, Joanna S Albala, Janghoo Lim, Carlene Fraughton, Estelle Llamosas, Sebiha Cevik, Camille Bex, Philippe Lamesch, Robert S Sikorski, Jean Vandenhaute, Huda Y Zoghbi, Alex Smolyar, Stephanie Bosak, Reynaldo Sequerra, Lynn Doucette-Stamm, Michael E Cusick, David E Hill, Frederick P Roth, and Marc Vidal. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437(7062):1173–1178, October 2005.

[68] Ulrich Stelzl, Uwe Worm, Maciej Lalowski, Christian Haenig, Felix H Brembeck, Heike Goehler, Martin Stroedicke, Martina Zenkner, Anke Schoenherr, Susanne Koeppen, Jan Timm, Sascha Mintzlaff, Claudia Abraham, Nicole Bock, Silvia Kietzmann, Astrid Goedde, Engin Toksöz, Anja Droege, Sylvia Krobitsch, Bernhard Korn, Walter Birchmeier, Hans Lehrach, and Erich E Wanker. A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, 122(6):957–968, September 2005.

[69] Christian von Mering, Roland Krause, Berend Snel, Michael Cornell, Stephen G Oliver, Stanley Fields, and Peer Bork. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417(6887):399–403, May 2002.

[70] Einat Sprinzak, Shmuel Sattath, and Hanah Margalit. How reliable are experimental protein-protein interaction data? *J Mol Biol*, 327(5):919–923, April 2003.

[71] Eric J Deeds, Orr Ashenberg, and Eugene I Shakhnovich. A simple physical model for scaling in protein-protein interaction networks. *Proceedings of the National Academy of Sciences of the United States of America*, 103(2):311–316, January 2006.

[72] Aled M Edwards, Bart Kus, Ronald Jansen, Dov Greenbaum, Jack Greenblatt, and Mark Gerstein. Bridging structural biology and genomics: assessing protein interaction data with known complexes. *Trends Genet*, 18(10):529–536, October 2002.

[73] Gary D Bader and Christopher W V Hogue. Analyzing yeast protein-protein interaction data obtained from different sources. *Nature biotechnology*, 20(10):991–997, October 2002.

[74] Haiyuan Yu, Pascal Braun, Muhammed A Yildirim, Irma Lemmens, Kavitha Venkatesan, Julie Sahalie, Tomoko Hirozane-Kishikawa, Fana Gebreab, Na Li, Nicolas Simonis, Tong Hao, Jean-François Rual, Amélie Dricot, Alexei Vazquez, Ryan R Murray, Christophe Simon, Leah Tardivo, Stanley Tam, Nenad Svrzikapa, Changyu Fan, Anne-Sophie de Smet, Adriana Motyl, Michael E Hudson, Juyong Park, Xiaofeng Xin, Michael E Cusick, Troy Moore, Charlie Boone, Michael Snyder, Frederick P Roth, Albert-Laszlo Barabasi, Jan Tavernier, David E Hill, and Marc Vidal. High-quality binary protein interaction map of the yeast interactome network. *Science (New York, NY)*, 322(5898):104–110, October 2008.

[75] Kavitha Venkatesan, Jean-François Rual, Alexei Vazquez, Ulrich Stelzl, Irma Lemmens, Tomoko Hirozane-Kishikawa, Tong Hao, Martina Zenkner, Xiaofeng Xin, Kwang-Il Goh, Muhammed A Yildirim, Nicolas Simonis, Kathrin Heinzmann, Fana Gebreab, Julie M Sahalie, Sebiha Cevik, Christophe Simon, Anne-Sophie de Smet, Elizabeth Dann, Alex Smolyar, Arunachalam Vinayagam, Haiyuan Yu, David Szeto, Heather Borick, Amélie Dricot, Niels Klitgord, Ryan R Murray, Chenwei Lin, Maciej Lalowski, Jan Timm, Kirstin Rau, Charles Boone,

Pascal Braun, Michael E Cusick, Frederick P Roth, David E Hill, Jan Tavernier, Erich E Wanker, Albert-Laszlo Barabasi, and Marc Vidal. An empirical framework for binary interactome mapping. *Nature methods*, 6(1):83–90, January 2009.

[76] Henning Hermjakob, Luisa Montecchi-Palazzi, Gary Bader, Jérôme Wojcik, Lukasz Salwinski, Arnaud Ceol, Susan Moore, Sandra Orchard, Ugis Sarkans, Christian von Mering, Bernd Roechert, Sylvain Poux, Eva Jung, Henning Mersch, Paul Kersey, Michael Lappe, Yixue Li, Rong Zeng, Debashis Rana, Macha Nikolski, Holger Husi, Christine Brun, K Shanker, Seth G N Grant, Chris Sander, Peer Bork, Weimin Zhu, Akhilesh Pandey, Alvis Brazma, Bernard Jacq, Marc Vidal, David Sherman, Pierre Legrain, Gianni Cesareni, Ioannis Xenarios, David Eisenberg, Boris Steipe, Chris Hogue, and Rolf Apweiler. The HUPO PSI's molecular interaction format–a community standard for the representation of protein interaction data. *Nature biotechnology*, 22(2):177–183, February 2004.

[77] Benjamin A Shoemaker and Anna R Panchenko. Deciphering protein-protein interactions. Part I. Experimental techniques and databases. *PLoS computational biology*, 3(3):e42, March 2007.

[78] Ioannis Xenarios, Lukasz Salwinski, Xiaoqun Joyce Duan, Patrick Higney, Sul-Min Kim, and David Eisenberg. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res*, 30(1):303–305, January 2002.

[79] Xiaoqun Joyce Duan, Ioannis Xenarios, and David Eisenberg. Describing biological protein interactions in terms of protein states and state transitions: the LiveDIP database. *Molecular & cellular proteomics : MCP*, 1(2):104–116, February 2002.

[80] G D Bader, I Donaldson, C Wolting, B F Ouellette, T Pawson, and C W Hogue. BIND–The Biomolecular Interaction Network Database. *Nucleic Acids Res*, 29(1):242–245, 2001.

[81] S Kerrien, Y Alam-Faruque, B Aranda, I Bancarz, A Bridge, C Derow, E Dimmer, M Feuermann, A Friedrichsen, R Huntley, C Kohler, J Khadake, C Leroy, A Liban, C Lieftink, L Montecchi-Palazzi, S Orchard, J Risse, K Robbe, B Roechert, D Thorneycroft, Y Zhang, R Apweiler, and H Hermjakob. IntAct–open source resource for molecular interaction data. *Nucleic Acids Res*, 35(Database issue):D561–5, 2007.

[82] Chris Stark, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, and Mike Tyers. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res*, 34 (Database issue):D535–9, January 2006.

[83] H Werner Mewes, Andreas Ruepp, Fabian Theis, Thomas Rattei, Mathias Walter, Dmitrij Frishman, Karsten Suhre, Manuel Spannagl, Klaus F X Mayer, Volker Stümpflen, and Alexey Antonov. MIPS: curated databases and comprehensive secondary data resources in 2010. *Nucleic Acids Res*, 39(Database issue):D220–4, January 2011.

[84] Ulrich Güldener, Martin Münsterkötter, Matthias Oesterheld, Philipp Pagel, Andreas Ruepp, Hans-Werner Mewes, and Volker Stümpflen. MPact: the MIPS protein interaction resource on yeast. *Nucleic Acids Res*, 34(Database issue):D436–41, January 2006.

[85] Johannes Goll, Seesandra V Rajagopala, Shen C Shiau, Hank Wu, Brian T Lamb, and Peter Uetz. MPIDB: the microbial protein interaction database. *Bioinformatics (Oxford, England)*, 24(15):1743–1744, August 2008.

[86] Suraj Peri, J Daniel Navarro, Ramars Amanchy, Troels Z Kristiansen, Chandra Kiran Jonnalagadda, Vineeth Surendranath, Vidya Niranjan, Babylakshmi Muthusamy, T K B Gandhi,

Mads Gronborg, Nieves Ibarrola, Nandan Deshpande, K Shanker, H N Shivashankar, B P Rashmi, M A Ramya, Zhixing Zhao, K N Chandrika, N Padma, H C Harsha, A J Yatish, M P Kavitha, Minal Menezes, Dipanwita Roy Choudhury, Shubha Suresh, Neelanjana Ghosh, R Saravana, Sreenath Chandran, Subhalakshmi Krishna, Mary Joy, Sanjeev K Anand, V Madavan, Ansamma Joseph, Guang W Wong, William P Schiemann, Stefan N Constantinescu, Lily Huang, Roya Khosravi-Far, Hanno Steen, Muneesh Tewari, Saghi Ghaffari, Gerard C Blobe, Chi V Dang, Joe G N Garcia, Jonathan Pevsner, Ole N Jensen, Peter Roepstorff, Krishna S Deshpande, Arul M Chinnaiyan, Ada Hamosh, Aravinda Chakravarti, and Akhilesh Pandey. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res*, 13(10):2363–2371, October 2003.

[87] Christian von Mering, Martijn Huynen, Daniel Jaeggi, Steffen Schmidt, Peer Bork, and Berend Snel. STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res*, 31(1):258–261, January 2003.

[88] Petras J Kundrotas and Emil Alexov. PROTCOM: searchable database of protein complexes enhanced with domain-domain structures. *Nucleic Acids Res*, 35(Database issue):D575–9, January 2007.

[89] Peter M Bowers, Matteo Pellegrini, Mike J Thompson, Joe Fierro, Todd O Yeates, and David Eisenberg. Prolinks: a database of protein functional linkages derived from coevolution. *Genome Biology*, 5(5):R35, 2004.

[90] Eduardo Andrés-León, Iakes Ezkurdia, Beatriz García, Alfonso Valencia, and David Juan. EcID. A database for the inference of functional interactions in E. coli. *Nucleic Acids Res*, 37 (Database issue):D629–35, January 2009.

[91] Benjamin A Shoemaker and Anna R Panchenko. Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners. *PLoS computational biology*, 3(4):e43, April 2007.

[92] A J Enright, I Iliopoulos, N C Kyrpides, and C A Ouzounis. Protein interaction maps for complete genomes based on gene fusion events. *Nature*, 402(6757):86–90, November 1999.

[93] E M Marcotte, M Pellegrini, M J Thompson, T O Yeates, and D Eisenberg. A combined algorithm for genome-wide prediction of protein function. *Nature*, 402(6757):83–86, November 1999.

[94] D M Burns, V Horn, J Paluh, and C Yanofsky. Evolution of the tryptophan synthetase of fungi. Analysis of experimentally fused Escherichia coli tryptophan synthetase alpha and beta chains. *J Biol Chem*, 265(4):2060–2069, February 1990.

[95] E M Marcotte, M Pellegrini, H L Ng, D W Rice, T O Yeates, and D Eisenberg. Detecting protein function and protein-protein interactions from genome sequences. *Science (New York, NY)*, 285(5428):751–753, July 1999.

[96] S Tsoka and C A Ouzounis. Prediction of protein interactions: metabolic enzymes are frequently involved in gene fusion. *Nature genetics*, 26(2):141–142, October 2000.

[97] Cynthia J Verjovsky Marcotte and Edward M Marcotte. Predicting functional linkages from gene fusions with confidence. *Applied bioinformatics*, 1(2):93–100, 2002.

[98] E Sprinzak and H Margalit. Correlated sequence-signatures as markers of protein-protein interaction. *J Mol Biol*, 311(4):681–692, August 2001.

[99] T Dandekar, B Snel, M Huynen, and P Bork. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends in biochemical sciences*, 23(9):324–328, September 1998.

[100] R Overbeek, M Fonstein, M D'Souza, G D Pusch, and N Maltsev. Use of contiguity on the chromosome to predict functional coupling. *In Silico Biol*, 1(2):93–108, 1999.

[101] M Huynen, B Snel, W 3rd Lathe, and P Bork. Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res*, 10(8):1204–1210, 2000.

[102] Sarah A Teichmann and M Madan Babu. Gene regulatory network growth by duplication. *Nature genetics*, 36(5):492–496, May 2004.

[103] E V Koonin, Y I Wolf, and L Aravind. Prediction of the archaeal exosome and its connections with the proteasome and the translation and transcription machineries by a comparative-genomic approach. *Genome Res*, 11(2):240–252, February 2001.

[104] Elena Evguenieva-Hackenberg, Pamela Walter, Elizabeth Hochleitner, Friedrich Lottspeich, and Gabriele Klug. An exosome-like complex in Sulfolobus solfataricus. *EMBO Rep*, 4(9): 889–893, September 2003.

[105] J G Lawrence and J R Roth. Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics*, 143(4):1843–1860, August 1996.

[106] T Gaasterland and M A Ragan. Microbial genescapes: phyletic and functional patterns of ORF distribution among prokaryotes. *Microbial & comparative genomics*, 3(4):199–217, 1998.

[107] M Pellegrini, Edward M Marcotte, M J Thompson, D Eisenberg, and T O Yeates. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A*, 96(8):4285–4288, 1999.

[108] Shailesh V Date and Edward M Marcotte. Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. *Nature biotechnology*, 21(9):1055–1062, 2003.

[109] Philipp Pagel, Philip Wong, and Dmitrij Frishman. A domain interaction map based on phylogenetic profiling. *J Mol Biol*, 344(5):1331–1346, December 2004.

[110] Juan A G Ranea, Daniel W A Buchan, Janet M Thornton, and Christine A Orengo. Evolution of protein superfamilies and bacterial genome size. *J Mol Biol*, 336(4):871–887, February 2004.

[111] Juan A G Ranea, Corin Yeats, Alastair Grant, and Christine A Orengo. Predicting protein function with hierarchical phylogenetic profiles: the Gene3D Phylo-Tuner method applied to eukaryotic genomes. *PLoS computational biology*, 3(11):e237, November 2007.

[112] Mitchell Levesque, Dennis Shasha, Wook Kim, Michael G Surette, and Philip N Benfey. Trait-to-gene: a computational method for predicting the function of uncharacterized genes. *Curr Biol*, 13(2):129–133, January 2003.

[113] Orland Gonzalez and Ralf Zimmer. Assigning functional linkages to proteins using phylogenetic profiles and continuous phenotypes. *Bioinformatics (Oxford, England)*, 24(10): 1257–1263, May 2008.

[114] Jie Wu, Simon Kasif, and Charles DeLisi. Identification of functional links between genes using phylogenetic profiles. *Bioinformatics (Oxford, England)*, 19(12):1524–1530, August 2003.

[115] Enrique Morett, Jan O Korbel, Emmanuvel Rajan, Gloria Saab-Rincon, Leticia Olvera, Maricela Olvera, Steffen Schmidt, Berend Snel, and Peer Bork. Systematic discovery of analogous enzymes in thiamin biosynthesis. *Nature biotechnology*, 21(7):790–795, July 2003.

[116] Jingchun Sun, Yixue Li, and Zhongming Zhao. Phylogenetic profiles for the prediction of protein-protein interactions: how to select reference organisms? *Biochemical and biophysical research communications*, 353(4):985–991, February 2007.

[117] Raja Jothi, Praveen F Cherukuri, Asba Tasneem, and Teresa M Przytycka. Co-evolutionary analysis of domains in interacting proteins reveals insights into domain-domain interactions mediating protein-protein interactions. *J Mol Biol*, 362(4):861–875, September 2006.

[118] Vijaykumar Yogesh Muley and Akash Ranjan. Effect of reference genome selection on the performance of computational methods for genome-wide protein-protein interaction prediction. *PloS one*, 7(7):e42057, 2012.

[119] Martin Simonsen, Stefan R Maetschke, and Mark A Ragan. Automatic selection of reference taxa for protein-protein interaction prediction with phylogenetic profiling. *Bioinformatics (Oxford, England)*, 28(6):851–857, March 2012.

[120] Daniel Barker and Mark Pagel. Predicting functional gene links from phylogenetic-statistical analyses of whole genomes. *PLoS computational biology*, 1(1):e3, June 2005.

[121] Ofir Cohen, Haim Ashkenazy, David Burstein, and Tal Pupko. Uncovering the co-evolutionary network among prokaryotic genes. *Bioinformatics (Oxford, England)*, 28(18):i389–i394, September 2012.

[122] Jean-Philippe Vert. A tree kernel to analyse phylogenetic profiles. *Bioinformatics (Oxford, England)*, 18 Suppl 1:S276–84, 2002.

[123] Yun Zhou, Rui Wang, Li Li, Xuefeng Xia, and Zhirong Sun. Inferring functional linkages between proteins from evolutionary scenarios. *J Mol Biol*, 359(4):1150–1159, June 2006.

[124] Yohan Kim and Shankar Subramaniam. Locally defined protein phylogenetic profiles reveal previously missed protein interactions and functional relationships. *Proteins*, 62(4):1115–1124, March 2006.

[125] Tamir Tuller, Yifat Felder, and Martin Kupiec. Discovering local patterns of co-evolution: computational aspects and biological examples. *BMC Bioinformatics*, 11:43, 2010.

[126] K J Fryxell. The coevolution of gene family trees. *Trends Genet*, 12(9):364–369, September 1996.

[127] R E van Kesteren, C P Tensen, A B Smit, J van Minnen, L F Kolakowski, W Meyerhof, D Richter, H van Heerikhuizen, E Vreugdenhil, and W P Geraerts. Co-evolution of ligand-receptor pairs in the vasopressin/oxytocin superfamily of bioactive peptides. *J Biol Chem*, 271(7):3619–3626, February 1996.

[128] S Pagès, A Bélaïch, J P Bélaïch, E Morag, R Lamed, Y Shoham, and E A Bayer. Species-specificity of the cohesin-dockerin interaction between Clostridium thermocellum and Clostridium cellulolyticum: prediction of specificity determinants of the dockerin domain. *Proteins*, 29(4):517–527, December 1997.

[129] C S Goh, A A Bogan, M Joachimiak, D Walther, and F E Cohen. Co-evolution of proteins with their interaction partners. *J Mol Biol*, 299(2):283–293, 2000.

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

[131] Chern-Sing Goh and Fred E Cohen. Co-evolutionary analysis reveals insights into protein-protein interactions. *J Mol Biol*, 324(1):177–192, November 2002.

[132] Florencio Pazos, Juan A G Ranea, David Juan, and Michael J E Sternberg. Assessing protein co-evolution in the context of the tree of life assists in the prediction of the interactome. *J Mol Biol*, 352(4):1002–1015, September 2005.

[133] Alessandra Devoto, H Andreas Hartmann, Pietro Piffanelli, Candace Elliott, Carl Simmons, Graziana Taramino, Chern-Sing Goh, Fred E Cohen, Brent C Emerson, Paul Schulze-Lefert, and Ralph Panstruga. Molecular phylogeny and evolution of the plant-specific seven-transmembrane MLO family. *Journal of molecular evolution*, 56(1):77–88, January 2003.

[134] Bernard Labedan, Ying Xu, Daniil G Naumoff, and Nicolas Glansdorff. Using quaternary structures to assess the evolutionary history of proteins: the case of the aspartate carbamoyltransferase. *Molecular biology and evolution*, 21(2):364–373, February 2004.

[135] Tonghai Dou, Chaoneng Ji, Shaohua Gu, Jiaxi Xu, Jian Xu, Kang Ying, Yi Xie, and Yumin Mao. Co-evolutionary analysis of insulin/insulin like growth factor 1 signal pathway in vertebrate species. *Frontiers in bioscience : a journal and virtual library*, 11:380–388, 2006.

[136] John M McPartland, Ryan W Norris, and C William Kilpatrick. Coevolution between cannabinoid receptors and endocannabinoid ligands. *Gene*, 397(1-2):126–135, August 2007.

[137] Zefeng Yang, Yong Zhou, Xuefeng Wang, Shiliang Gu, Jianmin Yu, Guohua Liang, Changjie Yan, and Chenwu Xu. Genomewide comparative phylogenetic and molecular evolutionary analysis of tubby-like protein family in Arabidopsis, rice, and poplar. *Genomics*, 92(4): 246–253, October 2008.

[138] Janet Piñero González, Olimpia Carrillo Farnés, Ana Tereza R Vasconcelos, and Abel González Pérez. Conservation of key members in the course of the evolution of the insulin signaling pathway. *Bio Systems*, 95(1):7–16, January 2009.

[139] Guy Leonard, Darren M Soanes, and Jamie R Stevens. Resolving the question of trypanosome monophyly: a comparative genomics approach using whole genome data sets with low taxon sampling. *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases*, 11(5):955–959, July 2011.

[140] David de Juan, Florencio Pazos, and Alfonso Valencia. Emerging methods in protein co-evolution. *Nat Rev Genet*, pages –, March 2013.

[141] Florencio Pazos and Alfonso Valencia. Protein co-evolution, co-adaptation and interactions. *The EMBO journal*, 27(20):2648–2655, October 2008.

[142] Leonid A Sazanov and Philip Hinchliffe. Structure of the hydrophilic domain of respiratory complex I from Thermus thermophilus. *Science (New York, NY)*, 311(5766):1430–1436, March 2006.

[143] Rachel S Edgar, Edward W Green, Yuwei Zhao, Gerben van Ooijen, Maria Olmedo, Ximing Qin, Yao Xu, Min Pan, Utham K Valekunja, Kevin A Feeney, Elizabeth S Maywood, Michael H Hastings, Nitin S Baliga, Martha Merrow, Andrew J Millar, Carl H Johnson, Charalambos P Kyriacou, John S O'Neill, and Akhilesh B Reddy. Peroxiredoxins are conserved markers of circadian rhythms. *Nature*, 485(7399):459–464, May 2012.

[144] Soon-Heng Tan, Zhuo Zhang, and See-Kiong Ng. ADVICE: Automated Detection and Validation of Interaction by Co-Evolution. *Nucleic Acids Res*, 32(Web Server issue):W69–72, July 2004.

[145] C Darwin. On the contrivances by which British and foreign orchids are fertilized by insects, and on the good effects of intercrossing. *John Murray*, page 365, 1862.

[146] P R Ehrlich and P H Raven. Butterflies and Plants: A Study in coevolution. *Evolution*, 18 (4):586–608, December 1964.

[147] J N Thompson. The coevolutionary process. *Chicago, Illinois, USA*, 1994.

[148] Andrés Moya, Juli Peretó, Rosario Gil, and Amparo Latorre. Learning how to live together: genomic insights into prokaryote-animal symbioses. *Nat Rev Genet*, 9(3):218–229, March 2008.

[149] A R Stone, D L Hawksworth, and Systematics Association. Coevolution and systematics. 1986.

[150] M S MS Hafner and S A SA Nadler. Phylogenetic trees support the coevolution of parasites and their hosts. *Nature*, 332(6161):258–259, March 1988.

[151] D J Futuyma. The uses of evolutionary biology. In *Science (New York, N.Y.)*, pages 41–42, January 1995.

[152] Leigh Van Valen. A new evolutionary law. *Evolutionary Theory*, 1:1–30–30, 1973.

[153] Leigh Van Valen. The red queen. *The American Naturalist*, 111(980):809–810, 1977.

[154] Leigh Van Valen. The red queen lives. *Nature*, 260:575, 1976.

[155] Theodosius Dobzhansky. Genetics of Natural Populations. Xix. Origin of Heterosis through Natural Selection in Populations of Drosophila Pseudoobscura. *Genetics*, 35(3):288, May 1950.

[156] Theodosius Grigorievich Dobzhansky. Genetics of the evolutionary process. *Columbia University Press*, 139, 1970.

[157] Bruce Wallace. On Coadaptation in Drosophila. *The American Naturalist*, 87:343–358, 1953.

[158] Bruce Wallace. Coadaptation Revisited. *Journal of Heredity*, 1991.

[159] W R Atchley, K R Wollenberg, W M Fitch, W Terhalle, and A W Dress. Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis. *Molecular biology and evolution*, 17(1):164–178, January 2000.

[160] 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 (Oxford, England)*, 21(17): 3482–3489, September 2005.

[161] Basant Tiwary, Besant K Tiwary, and Wen-Hsiung Li. Parallel evolution between aromatase and androgen receptor in the animal kingdom. *Molecular biology and evolution*, 26(1): 123–129, January 2009.

[162] Kwangbom Choi and Shawn M Gomez. Comparison of phylogenetic trees through alignment of embedded evolutionary distances. *BMC Bioinformatics*, 10:423, 2009.

[163] Tetsuya Sato, Yoshihiro Yamanishi, Katsuhisa Horimoto, Minoru Kanehisa, and Hiroyuki Toh. Partial correlation coefficient between distance matrices as a new indicator of protein-protein interactions. *Bioinformatics (Oxford, England)*, 22(20):2488–2492, October 2006.

[164] David Juan, Florencio Pazos, and Alfonso Valencia. High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. *Proc Natl Acad Sci U S A*, 105(3):934–939, January 2008.

[165] Elisabeth R M Tillier and Robert L Charlebois. The human protein coevolution network. *Genome research*, 19(10):1861–1871, October 2009.

[166] Alexandr Bezginov, Gregory W Clark, Robert L Charlebois, Vaqaar-Un-Nisa Dar, and Elisabeth R M Tillier. Coevolution reveals a network of human proteins originating with multicellularity. *Molecular biology and evolution*, 30(2):332–346, February 2013.

[167] Alex Rodionov, Alexandr Bezginov, Jonathan Rose, and Elisabeth RM Tillier. A new, fast algorithm for detecting protein coevolution using maximum compatible cliques. *Algorithms for molecular biology : AMB*, 6:17, 2011.

[168] Arun K Ramani and Edward M Marcotte. Exploiting the co-evolution of interacting proteins to discover interaction specificity. *J Mol Biol*, 327(1):273–284, March 2003.

[169] José M G Izarzugaza, David Juan, Carles Pons, Juan A G Ranea, Alfonso Valencia, and Florencio Pazos. TSEMA: interactive prediction of protein pairings between interacting families. *Nucleic Acids Res*, 34(Web Server issue):W315–9, July 2006.

[170] Elisabeth R M Tillier, Laurence Biro, Ginny Li, and Desiree Tillo. Codep: maximizing co-evolutionary interdependencies to discover interacting proteins. *Proteins*, 63(4):822–831, June 2006.

[171] José M G Izarzugaza, David Juan, Carles Pons, Florencio Pazos, and Alfonso Valencia. Enhancing the prediction of protein pairings between interacting families using orthology information. *BMC Bioinformatics*, 9:35, 2008.

[172] Iman Hajirasouliha, Alexander Schönhuth, David de Juan, Alfonso Valencia, and S Cenk Sahinalp. Mirroring co-evolving trees in the light of their topologies. *Bioinformatics (Oxford, England)*, 28(9):1202–1208, May 2012.

[173] Patrick Aloy and Robert B Russell. Structural systems biology: modelling protein interactions. *Nature reviews Molecular cell biology*, 7(3):188–197, March 2006.

[174] Maricel G Kann, Raja Jothi, Praveen F Cherukuri, and Teresa M Przytycka. Predicting protein domain interactions from coevolution of conserved regions. *Proteins*, 67(4):811–820, June 2007.

[175] Julian Mintseris and Zhiping Weng. Structure, function, and evolution of transient and obligate protein-protein interactions. *Proceedings of the National Academy of Sciences of the United States of America*, 102(31):10930–10935, August 2005.

[176] Maricel G Kann, Benjamin A Shoemaker, Anna R Panchenko, and Teresa M Przytycka. Correlated evolution of interacting proteins: looking behind the mirrortree. *J Mol Biol*, 385(1): 91–98, January 2009.

[177] Luke Hakes, Simon C Lovell, Stephen G Oliver, and David L Robertson. Specificity in protein interactions and its relationship with sequence diversity and coevolution. *Proceedings of the National Academy of Sciences of the United States of America*, 104(19):7999–8004, May 2007.

[178] Hunter B Fraser, Aaron E Hirsh, Lars M Steinmetz, Curt Scharfe, and Marcus W Feldman. Evolutionary rate in the protein interaction network. *Science (New York, NY)*, 296(5568): 750–752, April 2002.

[179] Nathaniel L Clark and Charles F Aquadro. A novel method to detect proteins evolving at correlated rates: identifying new functional relationships between coevolving proteins. *Molecular biology and evolution*, 27(5):1152–1161, May 2010.

[180] Nathan L Clark, Eric Alani, and Charles F Aquadro. Evolutionary rate covariation reveals shared functionality and coexpression of genes. *Genome research*, 22(4):714–720, April 2012.

[181] M B Eisen, P T Spellman, P O Brown, and D Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, 95(25):14863–14868, December 1998.

[182] C Pal, B Papp, and L D Hurst. Highly expressed genes in yeast evolve slowly. *Genetics*, 158 (2):927–931, June 2001.

[183] Hunter B Fraser, Aaron E Hirsh, Dennis P Wall, and Michael B Eisen. Coevolution of gene expression among interacting proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 101(24):9033–9038, June 2004.

[184] Sankar Subramanian and Sudhir Kumar. Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. *Genetics*, 168(1):373–381, September 2004.

[185] Yiwen Chen and Nikolay V Dokholyan. The coordinated evolution of yeast proteins is constrained by functional modularity. *Trends Genet*, 22(8):416–419, August 2006.

[186] D Allan Drummond, Alpan Raval, and Claus O Wilke. A single determinant dominates the rate of yeast protein evolution. *Molecular biology and evolution*, 23(2):327–337, February 2006.

[187] Zhi Liang, Meng Xu, Maikun Teng, Liwen Niu, and Jiarui Wu. Coevolution is a short-distance force at the protein interaction level and correlates with the modular organization of protein networks. *FEBS letters*, 584(19):4237–4240, October 2010.

[188] Joel R Bock and David A Gough. Whole-proteome interaction mining. *Bioinformatics (Oxford, England)*, 19(1):125–134, January 2003.

[189] Y Yamanishi, J-P Vert, and M Kanehisa. Protein network inference from multiple genomic data: a supervised approach. *Bioinformatics (Oxford, England)*, 20 Suppl 1:i363–70, August 2004.

[190] Shawn Martin, Diana Roe, and Jean-Loup Faulon. Predicting protein-protein interactions using signature products. *Bioinformatics (Oxford, England)*, 21(2):218–226, January 2005.

[191] Roger A Craig and Li Liao. Phylogenetic tree information aids supervised learning for predicting protein-protein interaction based on distance matrices. *BMC Bioinformatics*, 8:6, 2007.

[192] Roger A Craig and Li Liao. Improving protein protein interaction prediction based on phylogenetic information using a least-squares support vector machine. *Annals of the New York Academy of Sciences*, 1115:154–167, December 2007.

[193] Nazar Zaki, Sanja Lazarova-Molnar, Wassim El-Hajj, and Piers Campbell. Protein-protein interaction based on pairwise similarity. *BMC Bioinformatics*, 10:150, 2009.

[194] Charlotte M Deane, Łukasz Salwiński, Ioannis Xenarios, and David Eisenberg. Protein interactions: two methods for assessment of the reliability of high throughput observations. *Molecular & cellular proteomics : MCP*, 1(5):349–356, May 2002.

[195] Ramazan Saeed and Charlotte M Deane. Protein protein interactions, evolutionary rate, abundance and age. *BMC Bioinformatics*, 7:128, 2006.

[196] Minghua Deng, Fengzhu Sun, and Ting Chen. Assessment of the reliability of protein-protein interactions and protein function prediction. *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing*, pages 140–151, 2003.

[197] Joel S Bader, Amitabha Chaudhuri, Jonathan M Rothberg, and John Chant. Gaining confidence in high-throughput protein interaction networks. *Nature biotechnology*, 22(1):78–85, January 2004.

[198] Roded Sharan, Silpa Suthram, Ryan M Kelley, Tanja Kuhn, Scott McCuine, Peter Uetz, Taylor Sittler, Richard M Karp, and Trey Ideker. Conserved patterns of protein interaction in multiple species. *Proceedings of the National Academy of Sciences of the United States of America*, 102(6):1974–1979, February 2005.

[199] Xiaotong Lin, Mei Liu, and Xue-wen Chen. Assessing reliability of protein-protein interactions by integrative analysis of data in model organisms. *BMC Bioinformatics*, 10 Suppl 4: S5, 2009.

[200] Ramazan Saeed and Charlotte Deane. An assessment of the uses of homologous interactions. *Bioinformatics (Oxford, England)*, 24(5):689–695, March 2008.

[201] Sven Mika and Burkhard Rost. Protein-protein interactions more conserved within species than across species. *PLoS computational biology*, 2(7):e79, July 2006.

[202] Rintaro Saito, Harukazu Suzuki, and Yoshihide Hayashizaki. Interaction generality, a measurement to assess the reliability of a protein-protein interaction. *Nucleic Acids Res*, 30(5): 1163–1168, March 2002.

[203] Rintaro Saito, Harukazu Suzuki, and Yoshihide Hayashizaki. Construction of reliable protein-protein interaction networks with a new interaction generality measure. *Bioinformatics (Oxford, England)*, 19(6):756–763, April 2003.

[204] Jin Chen, Wynne Hsu, Mong Li Lee, and See-Kiong Ng. Increasing confidence of protein interactomes using network topological metrics. *Bioinformatics (Oxford, England)*, 22(16): 1998–2004, August 2006.

[205] Tom Fawcett. An introduction to ROC analysis. 27(8):861–874, June 2006.

[206] Pall F Jonsson and Paul A Bates. Global topological features of cancer proteins in the human interactome. *Bioinformatics (Oxford, England)*, 22(18):2291–2297, September 2006.

[207] Ashwini Patil and Haruki Nakamura. Filtering high-throughput protein-protein interaction data using a combination of genomic features. *BMC Bioinformatics*, 6:100, 2005.

[208] Asa Ben-Hur and William Stafford Noble. Choosing negative examples for the prediction of protein-protein interactions. *BMC Bioinformatics*, 7 Suppl 1:S2, 2006.

[209] Pawel Smialowski, Philipp Pagel, Philip Wong, Barbara Brauner, Irmtraud Dunger, Gisela Fobo, Goar Frishman, Corinna Montrone, Thomas Rattei, Dmitrij Frishman, and Andreas Ruepp. The Negatome database: a reference set of non-interacting protein pairs. *Nucleic Acids Res*, 38(Database issue):D540–4, January 2010.

[210] David Juan, Florencio Pazos, and Alfonso Valencia. Co-evolution and co-adaptation in protein networks. *FEBS Lett*, 582(8):1225–1230, 2008.

[211] Ingrid M Keseler, Julio Collado-Vides, Socorro Gama-Castro, John Ingraham, Suzanne Paley, Ian T Paulsen, Martin Peralta-Gil, and Peter D Karp. EcoCyc: a comprehensive database resource for Escherichia coli. *Nucleic Acids Res*, 33(Database issue):D334–7, January 2005.

[212] UniProt Consortium. The Universal Protein Resource (UniProt). *Nucleic Acids Res*, 35 (Database issue):D193–7, January 2007.

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

[214] Paul Kersey, Lawrence Bower, Lorna Morris, Alan Horne, Robert Petryszak, Carola Kanz, Alexander Kanapin, Ujjwal Das, Karine Michoud, Isabelle Phan, Alexandre Gattiker, Tamara Kulikova, Nadeem Faruque, Karyn Duggan, Peter Mclaren, Britt Reimholz, Laurent Duret, Simon Penel, Ingmar Reuter, and Rolf Apweiler. Integr8 and Genome Reviews: integrated views of complete genomes and proteomes. *Nucleic Acids Res*, 33(Database issue):D297–D302, December 2004.

[215] Robert C Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, 32(5):1792–1797, 2004.

[216] Ramu Chenna, Hideaki Sugawara, Tadashi Koike, Rodrigo Lopez, Toby J Gibson, Desmond G Higgins, and Julie D Thompson. Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res*, 31(13):3497–3500, July 2003.

[217] Burkhard Rost. PHD: Predicting one-dimensional protein structure by profile-based neural networks. In Russell F Doolittle, editor, *Methods in Enzymology*, pages 525–539. Academic Press, 1996.

[218] M A Larkin, G Blackshields, N P Brown, R Chenna, P A McGettigan, H McWilliam, F Valentin, I M Wallace, A Wilm, R Lopez, J D Thompson, T J Gibson, and D G Higgins. Clustal W and Clustal X version 2.0. *Bioinformatics (Oxford, England)*, 23(21):2947–2948, November 2007.

[219] Dennis A Benson, Ilene Karsch-Mizrachi, David J Lipman, James Ostell, and Eric W Sayers. GenBank. *Nucleic Acids Res*, 37(Database issue):D26–31, January 2009.

[220] Minoru Kanehisa, Susumu Goto, Shuichi Kawashima, Yasushi Okuno, and Masahiro Hattori. The KEGG resource for deciphering the genome. *Nucleic Acids Res*, 32(Database issue): D277–80, 2004.

[221] Albert J Vilella, Jessica Severin, Abel Ureta-Vidal, Li Heng, Richard Durbin, and Ewan Birney. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res*, 19(2):327–335, February 2009.

[222] D M Raskin and P A de Boer. MinDE-dependent pole-to-pole oscillation of division inhibitor MinC in Escherichia coli. *J Bacteriol*, 181(20):6419–6424, October 1999.

[223] P M Jones and A M George. The ABC transporter structure and mechanism: perspectives on recent research. *Cellular and molecular life sciences : CMLS*, 61(6):682–699, March 2004.

[224] Zhi Wang and Jianzhi Zhang. Why is the correlation between gene importance and gene evolutionary rate so weak? *PLoS Genet*, 5(1):e1000329, January 2009.

[225] B Rost and C Sander. Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins*, 19(1):55–72, May 1994.

[226] Dariusz Przybylski and Burkhard Rost. Alignments grow, secondary structure prediction improves. *Proteins*, 46(2):197–205, February 2002.

[227] Simon C Lovell and David L Robertson. An integrated view of molecular coevolution in protein-protein interactions. *Molecular biology and evolution*, 27(11):2567–2575, November 2010.

[228] Chong Shou, Nitin Bhardwaj, Hugo Y K Lam, Koon-Kiu Yan, Philip M Kim, Michael Snyder, and Mark B Gerstein. Measuring the evolutionary rewiring of biological networks. *PLoS computational biology*, 7(1):e1001050, 2011.

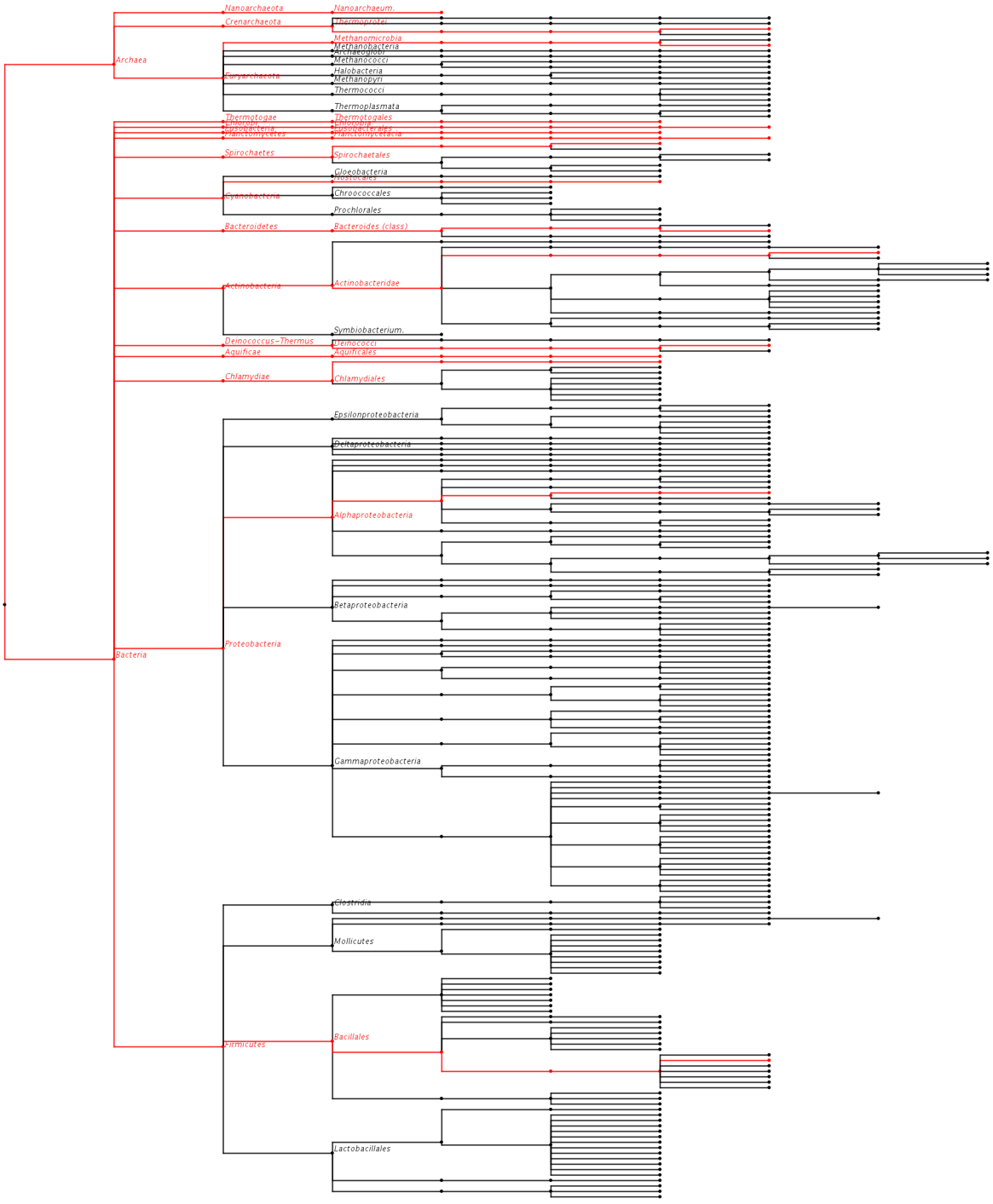
[229] Patrick Aloy, Hugo Ceulemans, Alexander Stark, and Robert B Russell. The relationship between sequence and interaction divergence in proteins. *J Mol Biol*, 332(5):989–998, October 2003.

[230] Raja Jothi, Teresa M Przytycka, and L Aravind. Discovering functional linkages and uncharacterized cellular pathways using phylogenetic profile comparisons: a comprehensive assessment. *BMC Bioinformatics*, 8:173, 2007.

Appendices

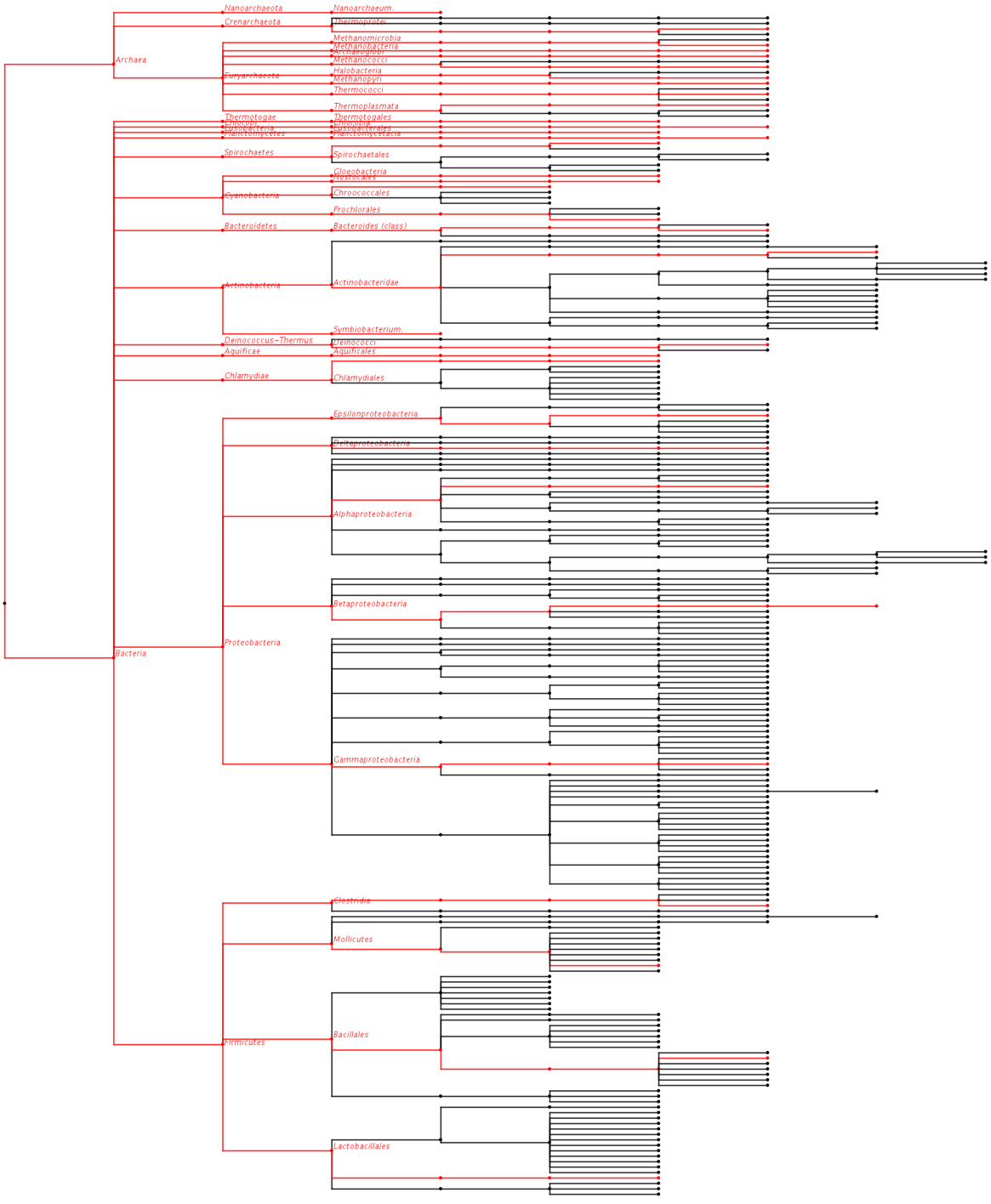
Sets of reference organisms

The next pages contain the list of organisms in the subsets created by two different taxonomic criteria: “nearest” - going from the reference organism to the root taking all the organisms in the resulting taxa - and “level” - the tree is successively cut and one organism is taken from each one of the resulting groups. A representation of whole taxonomic tree (black) with the selected organisms (red) is also included in the next pages.


Archaea
Nanoarchaeota
Crenarchaeota
Euryarchaeota
Thermotogae
Chlorobi
Cytophagales
Planctomycetes
Spirochaetes
Cyanobacteria
Bacteroidetes
Actinobacteria
Deinococcus–Thermus
Aquificae
Chlamydiae
Bacteria
Proteobacteria
Firmicutes
Nanoarchaeum
Thermoprotei
Methanomicrobia
Methanobacteria
Archaeoglobi
Methanococci
Halobacteria
Methanopyri
Thermococci
Thermoplasmata
Thermotogales
Chlorobia
Cytophagales
Planctomycetacia
Spirochaetales
Gloeobacteria
Nostocales
Chroococcales
Prochlorales
Bacteroides (class)
Actinobacteridae
Symbiobacterium
Deinococci
Aquificales
Chlamydiales
Epsilonproteobacteria
Deltaproteobacteria
Alphaproteobacteria
Betaproteobacteria
Gammaproteobacteria
Clostridia
Mollicutes
Bacillales
Lactobacillales

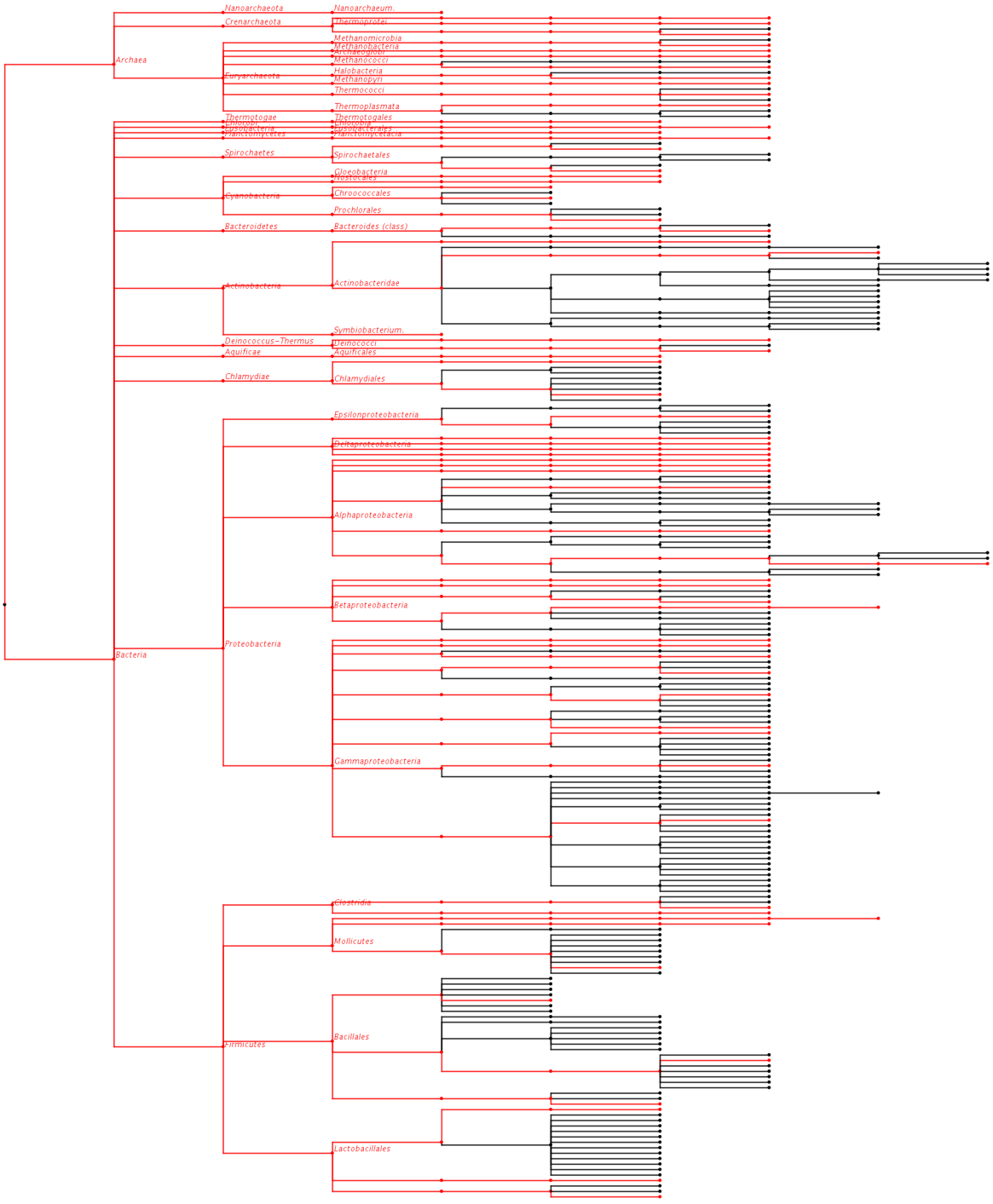
Level 2

Anabaena sp.
Aquifex aeolicus
Bacillus cereus ATCC 10987
Bacteroides thetaiotaomicron
Bradyrhizobium japonicum
Chlorobium tepidum
Fusobacterium nucleatum
Leptospira interrogans lai
Methanosarcina acetivorans
Nanoarchaeum equitans
Parachlamydia sp.
Rhodopirellula baltica
Streptomyces coelicolor
Sulfolobus solfataricus
Thermotoga maritima
Thermus thermophilus HB8


Archaea
Nanoarchaeota
Crenarchaeota
Euryarchaeota
Thermotogae
Spirochaetes
Cyanobacteria
Bacteroidetes
Actinobacteria
Deinococcus–Thermus
Aquificae
Chlamydiae
Bacteria
Proteobacteria
Firmicutes
Nanoarchaeum
Thermoprotei
Methanomicrobia
Methanobacteria
Methanococci
Halobacteria
Methanopyri
Thermococci
Thermoplasmata
Thermotogales
Spirochaetales
Gloeobacteria
Chroococcales
Prochlorales
Bacteroides (class)
Actinobacteridae
Symbiobacterium
Deinococci
Aquificales
Chlamydiales
Epsilonproteobacteria
Deltaproteobacteria
Alphaproteobacteria
Betaproteobacteria
Gammaproteobacteria
Clostridia
Mollicutes
Bacillales
Lactobacillales

Level 3

Anabaena sp.
Aquifex aeolicus
Archaeoglobus fulgidus
Bacillus cereus ATCC 10987
Bacteroides thetaiotaomicron
Burkholderia pseudomallei
Chlorobium tepidum
Clostridium acetobutylicum
Desulfovibrio vulgaris
Enterococcus faecalis
Fusobacterium nucleatum
Gloeobacter violaceus
Haloarcula marismortui
Leptospira interrogans lai
Methanobacterium thermoautotrophicum
Methanococcus maripaludis
Methanopyrus kandleri
Methanosarcina acetivorans
Mycoplasma penetrans
Nanoarchaeum equitans
Parachlamydia sp.
Picrophilus torridus
Prochlorococcus marinus MIT 9313
Pseudomonas aeruginosa
Pyrococcus horikoshii
Rhizobium loti
Rhodopirellula baltica
Streptomyces coelicolor
Sulfolobus solfataricus
Symbiobacterium thermophilum
Synechocystis sp.
Thermotoga maritima
Thermus thermophilus HB8
Wolinella succinogenes


Archaea
Nanoarchaeota
Crenarchaeota
Euryarchaeota
Thermotogae
Spirochaetes
Cyanobacteria
Bacteroidetes
Actinobacteria
Deinococcus–Thermus
Aquificae
Chlamydiae
Bacteria
Proteobacteria
Firmicutes
Nanoarchaeum
Thermoprotei
Methanomicrobia
Methanobacteria
Archaeoglobi
Methanococci
Halobacteria
Methanopyri
Thermococci
Thermoplasmata
Thermotogales
Spirochaetales
Gloeobacteria
Chroococcales
Prochlorales
Bacteroides (class)
Actinobacteridae
Symbiobacterium
Deinococci
Aquificales
Chlamydiales
Epsilonproteobacteria
Deltaproteobacteria
Alphaproteobacteria
Betaproteobacteria
Gammaproteobacteria
Clostridia
Mollicutes
Bacillales
Lactobacillales

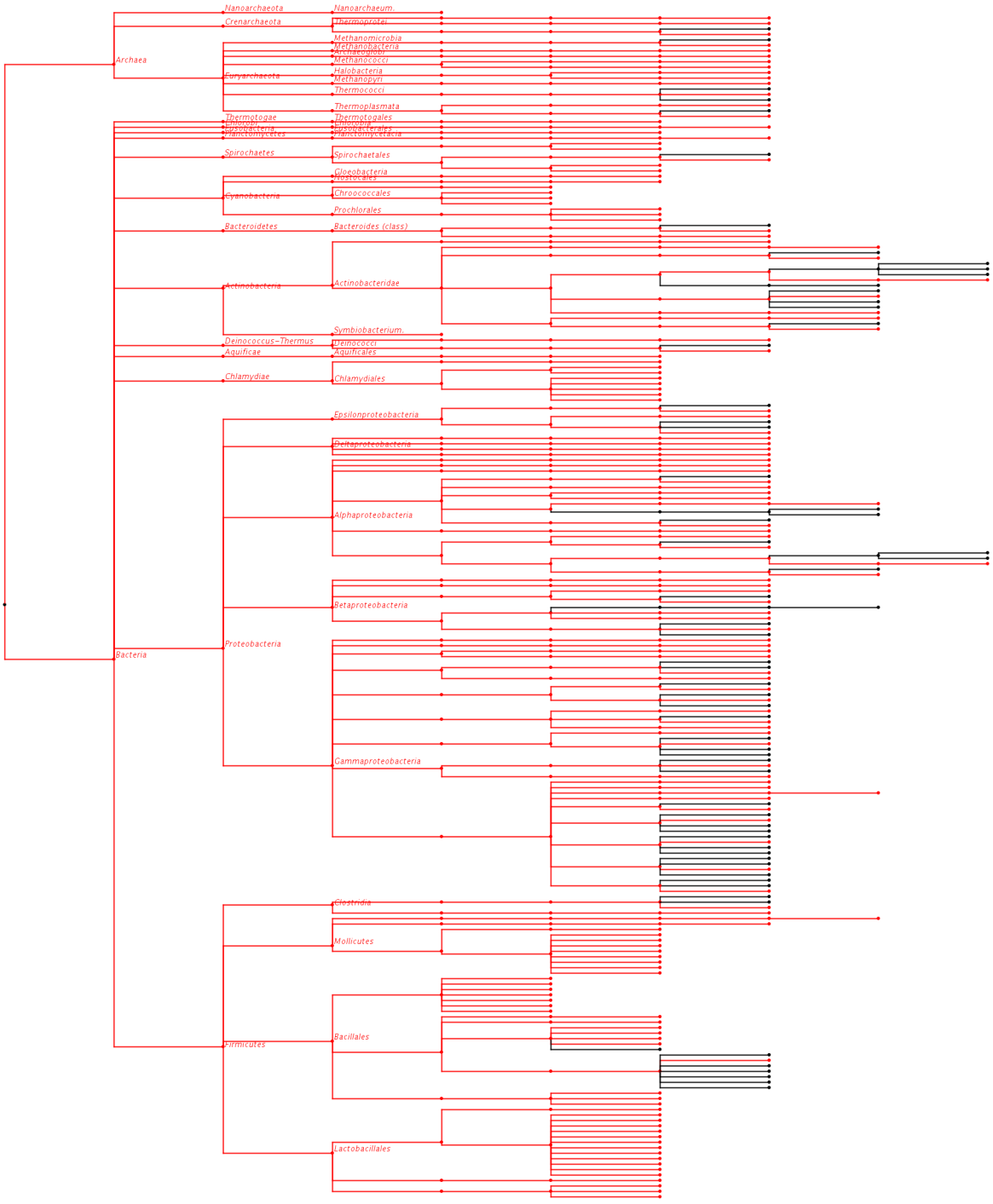
Level 4

Aeropyrum pernix
Anabaena sp.
Aquifex aeolicus
Archaeoglobus fulgidus
Azoarcus sp.
Bacillus cereus ATCC 10987
Bacteroides thetaiotaomicron
Bdellovibrio bacteriovorus
Bifidobacterium longum
Burkholderia pseudomallei
Caulobacter crescentus
Chlamydia pneumoniae TW-183
Chlorobium tepidum
Clostridium acetobutylicum
Deinococcus radiodurans
Desulfotalea psychrophila
Desulfovibrio vulgaris
Enterococcus faecalis
Escherichia coli O6 UPEC
Francisella tularensis
Fusobacterium nucleatum
Geobacter sulfurreducens
Gloeobacter violaceus
Gluconobacter oxydans
Haloarcula marismortui
Lactobacillus acidophilus
Lactococcus lactis
Legionella pneumophila Philadelphia 1
Leptospira interrogans Icterohaemorrhagiae
Listeria innocua
Mesoplasma florum
Methanobacterium thermoautotrophicum
Methanococcus maripaludis
Methanopyrus kandleri
Methanosarcina acetivorans
Methylococcus capsulatus
Mycoplasma penetrans
Nanoarchaeum equitans
Neisseria meningitidis A
Nitrosomonas europaea
Onion yellows phytoplasma
Parachlamydia sp.
Pasteurella multocida
Photobacterium profundum
Picrophilus torridus
Prochlorococcus marinus MIT 9313
Pseudomonas aeruginosa
Pyrobaculum aerophilum
Pyrococcus horikoshii
Rhizobium loti
Rhodopirellula baltica
Rickettsia conorii
Shewanella oneidensis
Silicibacter pomeroyi
Staphylococcus aureus Mu50
Streptomyces coelicolor
Sulfolobus tokodaii
Symbiobacterium thermophilum
Synechococcus sp. PCC 6301
Synechocystis sp.
Thermoanaerobacter tengcongensis
Thermotoga maritima
Thermus thermophilus HB27
Treponema denticola
Wolinella succinogenes
Xanthomonas oryzae
Zymomonas mobilis


Archaea
Nanoarchaeota
Crenarchaeota
Euryarchaeota
Nanoarchaeum
Thermoprotei
Methanomicrobia
Methanobacteria
Archaeoglobi
Methanococci
Halobacteria
Methanopyri
Thermococci
Thermoplasmata
Thermotogae
Chloroflexi
Planctomycetes
Thermotogales
Chloroflexi
Sphaerobacterales
Planctomycetacia
Spirochaetes
Spirochaetales
Gloeobacteria
Nostocales
Cyanobacteria
Chroococcales
Prochlorales
Bacteroidetes
Bacteroides (class)
Actinobacteria
Actinobacteridae
Deinococcus–Thermus
Aquificae
Symbiobacterium
Thermococci
Aquificales
Chlamydiae
Chlamydiales
Bacteria
Epsilonproteobacteria
Deltaproteobacteria
Alphaproteobacteria
Betaproteobacteria
Proteobacteria
Gammaproteobacteria
Clostridia
Mollicutes
Bacillales
Firmicutes
Lactobacillales

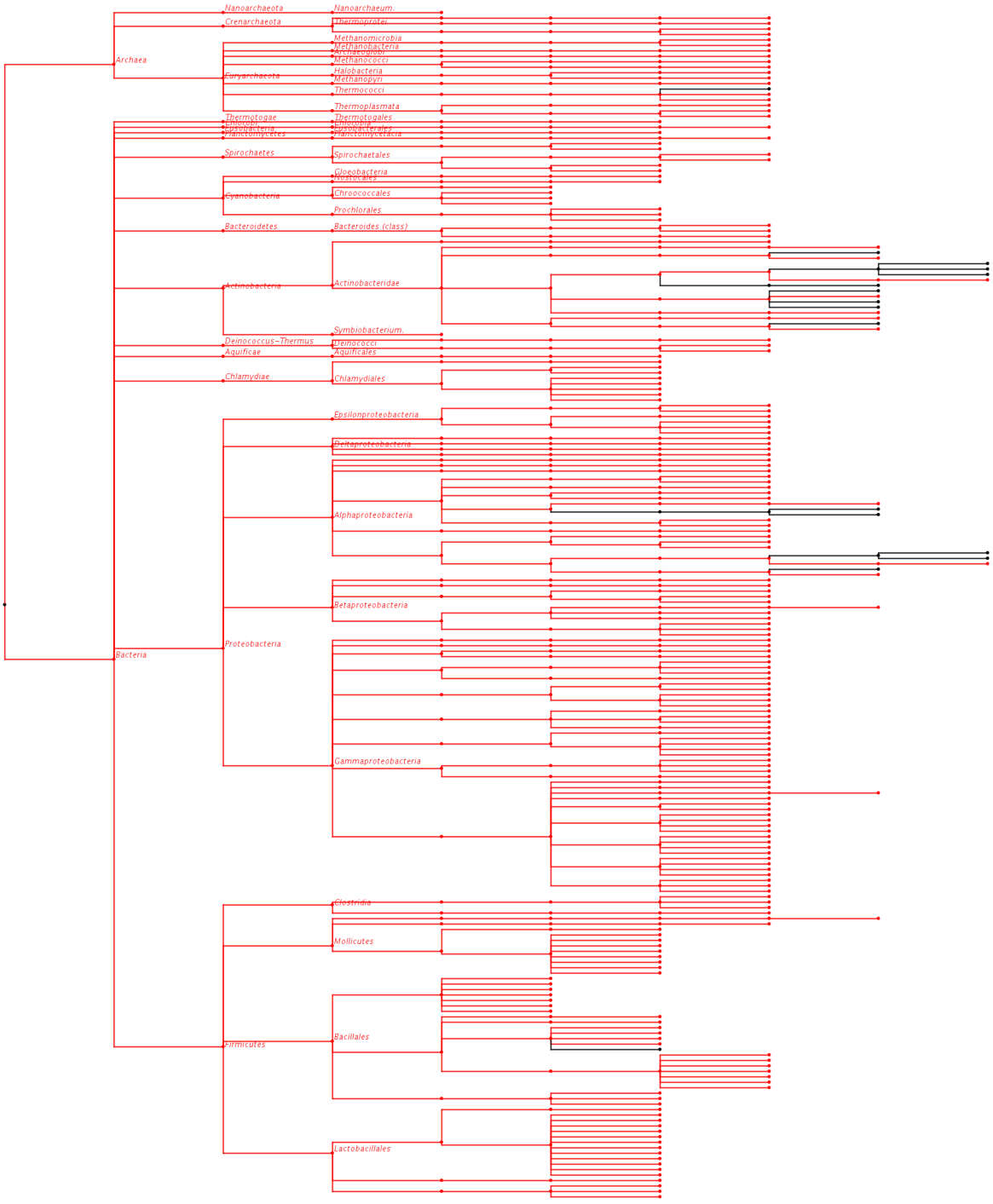
Level 5

Acinetobacter sp.
Aeropyrum pernix
Anabaena sp.
Aquifex aeolicus
Archaeoglobus fulgidus
Azoarcus sp.
Bacillus cereus ATCC 10987
Bacteroides thetaiotaomicron
Bartonella henselae
Bdellovibrio bacteriovorus
Bifidobacterium longum
Bordetella bronchiseptica
Borrelia garinii
Brucella melitensis
Campylobacter jejuni NCTC 11168
Caulobacter crescentus
Chlamydia muridarum
Chlamydia pneumoniae TW-183
Chlorobium tepidum
Clostridium acetobutylicum
Coxiella burnetii
Deinococcus radiodurans
Desulfotalea psychrophila
Desulfovibrio vulgaris
Ehrlichia ruminantium CIRAD
Enterococcus faecalis
Escherichia coli O6 UPEC
Francisella tularensis
Fusobacterium nucleatum
Geobacillus kaustophilus
Geobacter sulfurreducens
Gloeobacter violaceus
Gluconobacter oxydans
Haloarcula marismortui
Idiomarina loihiensis
Lactobacillus acidophilus
Lactococcus lactis
Legionella pneumophila Philadelphia 1
Leifsonia xyli
Leptospira interrogans Icterohaemorrhagiae
Listeria innocua
Mesoplasma florum
Methanobacterium thermoautotrophicum
Methanococcus jannaschii
Methanococcus maripaludis
Methanopyrus kandleri
Methanosarcina acetivorans
Methylococcus capsulatus
Mycoplasma penetrans
Nanoarchaeum equitans
Neisseria meningitidis A
Nitrosomonas europaea
Nocardia farcinica
Oceanobacillus iheyensis
Onion yellows phytoplasma
Parachlamydia sp.
Pasteurella multocida
Photobacterium profundum
Picrophilus torridus
Porphyromonas gingivalis
Prochlorococcus marinus MIT 9313
Propionibacterium acnes
Pseudomonas aeruginosa
Pyrobaculum aerophilum
Pyrococcus horikoshii
Ralstonia solanacearum
Rhizobium loti
Rhizobium meliloti
Rhodopirellula baltica
Rhodopseudomonas palustris
Rickettsia conorii
Shewanella oneidensis
Silicibacter pomeroyi
Staphylococcus aureus COL
Staphylococcus aureus MRSA252
Staphylococcus aureus MSSA476
Staphylococcus aureus MW2
Staphylococcus aureus Mu50
Staphylococcus aureus N315
Staphylococcus epidermidis ATCC 12228
Streptococcus agalactiae V
Streptomyces avermitilis
Sulfolobus tokodaii
Symbiobacterium thermophilum
Synechococcus elongatus
Synechococcus sp. PCC 6301
Synechococcus sp. WH8102
Synechocystis sp.
Thermoanaerobacter tengcongensis
Thermoplasma volcanium
Thermotoga maritima
Thermus thermophilus HB27
Treponema denticola
Ureaplasma parvum
Wolinella succinogenes
Xanthomonas oryzae
Zymomonas mobilis


Archaea
Nanoarchaeota
Crenarchaeota
Euryarchaeota
Nanoarchaeum
Thermoprotei
Methanomicrobia
Methanobacteria
Archaeoglobi
Methanococci
Halobacteria
Methanopyri
Thermococci
Thermoplasmata
Thermotogae
Chloroflexi
Planctomycetes
Thermotogales
Chloroflexi
Epsilonbacteraeles
Planctomycetacia
Spirochaetes
Spirochaetales
Gloeobacteria
Nostocales
Cyanobacteria
Chroococcales
Prochlorales
Bacteroidetes
Bacteroides (class)
Actinobacteria
Actinobacteridae
Deinococcus–Thermus
Aquificae
Symbiobacterium
Thermococci
Aquificales
Chlamydiae
Chlamydiales
Bacteria
Epsilonproteobacteria
Deltaproteobacteria
Alphaproteobacteria
Betaproteobacteria
Proteobacteria
Gammaproteobacteria
Clostridia
Mollicutes
Bacillales
Firmicutes
Lactobacillales

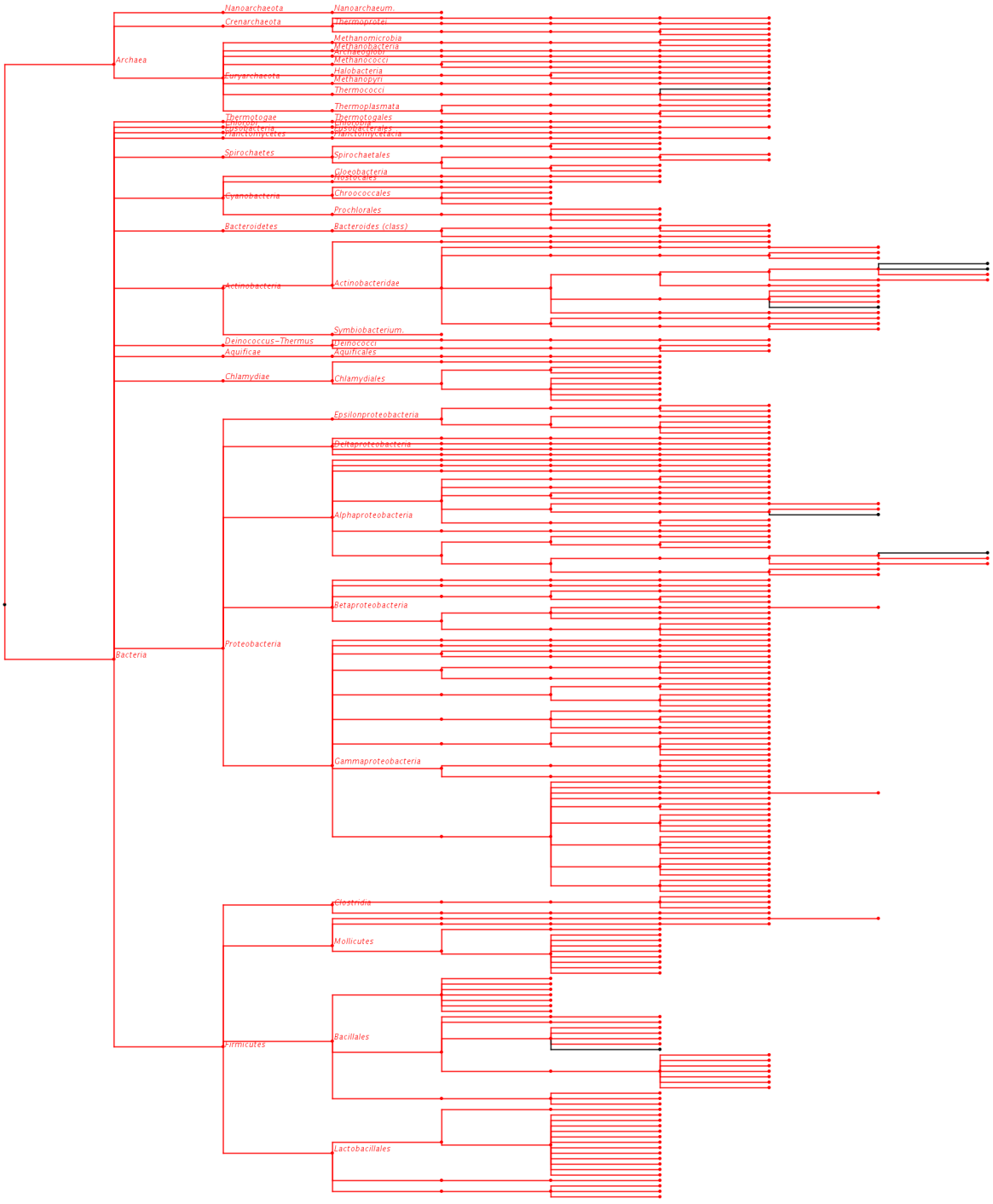
Level 6

Acinetobacter sp.
Aeropyrum pernix
Anabaena sp.
Anaplasma marginale
Aquifex aeolicus
Archaeoglobus fulgidus
Azoarcus sp.
Bacillus cereus ATCC 10987
Bacillus clausii
Bacillus halodurans
Bacillus licheniformis Goettingen
Bacillus subtilis
Bacteroides thetaiotaomicron
Bartonella henselae
Bdellovibrio bacteriovorus
Bifidobacterium longum
Bordetella bronchiseptica
Borrelia garinii
Bradyrhizobium japonicum
Brucella melitensis
Buchnera aphidicola Acyrthosiphon pisum
Burkholderia mallei
Campylobacter jejuni NCTC 11168
Candidatus Blochmannia floridanus
Caulobacter crescentus
Chlamydia muridarum
Chlamydia pneumoniae AR39
Chlamydia pneumoniae CWL029
Chlamydia pneumoniae J138
Chlamydia pneumoniae TW-183
Chlamydia trachomatis
Chlamydophila caviae
Chlorobium tepidum
Chromobacterium violaceum
Clostridium acetobutylicum
Corynebacterium glutamicum Nakagawa
Coxiella burnetii
Deinococcus radiodurans
Desulfotalea psychrophila
Desulfovibrio vulgaris
Ehrlichia ruminantium Gardel
Enterococcus faecalis
Erwinia carotovora
Escherichia coli O6 UPEC
Francisella tularensis
Fusobacterium nucleatum
Geobacillus kaustophilus
Geobacter sulfurreducens
Gloeobacter violaceus
Gluconobacter oxydans
Haemophilus ducreyi
Haloarcula marismortui
Halobacterium salinarium
Helicobacter pylori ATCC 700392
Idiomarina loihiensis
Lactobacillus acidophilus
Lactobacillus johnsonii
Lactobacillus plantarum
Lactococcus lactis
Legionella pneumophila Philadelphia 1
Leifsonia xyli
Leptospira interrogans Icterohaemorrhagiae
Leptospira interrogans lai
Listeria innocua
Listeria monocytogenes 1/2a
Listeria monocytogenes 4b
Mannheimia succiniciproducens
Mesoplasma florum
Methanobacterium thermoautotrophicum
Methanococcus jannaschii
Methanococcus maripaludis
Methanopyrus kandleri
Methanosarcina acetivorans
Methylococcus capsulatus
Mycobacterium paratuberculosis
Mycoplasma gallisepticum
Mycoplasma genitalium
Mycoplasma hyopneumoniae
Mycoplasma mobile
Mycoplasma mycoides
Mycoplasma penetrans
Mycoplasma pneumoniae
Mycoplasma pulmonis
Nanoarchaeum equitans
Neisseria meningitidis A
Nitrosomonas europaea
Nocardia farcinica
Oceanobacillus iheyensis
Onion yellows phytoplasma
Parachlamydia sp.
Pasteurella multocida
Photobacterium profundum
Photorhabdus luminescens
Picrophilus torridus
Porphyromonas gingivalis
Prochlorococcus marinus CCMP 1375
Prochlorococcus marinus CCMP 1378
Prochlorococcus marinus MIT 9313
Propionibacterium acnes
Pseudomonas aeruginosa
Pyrobaculum aerophilum
Pyrococcus horikoshii
Ralstonia solanacearum
Rhizobium loti
Rhizobium meliloti
Rhodopirellula baltica
Rhodopseudomonas palustris
Rickettsia conorii
Salmonella typhi CT18
Shewanella oneidensis
Shigella flexneri 301
Silicibacter pomeroyi
Staphylococcus aureus COL
Staphylococcus aureus MRSA252
Staphylococcus aureus MSSA476
Staphylococcus aureus MW2
Staphylococcus aureus Mu50
Staphylococcus aureus N315
Staphylococcus epidermidis ATCC 12228
Streptococcus agalactiae III
Streptococcus agalactiae V
Streptococcus mutans
Streptococcus pneumoniae ATCC BAA-255
Streptococcus pneumoniae TIGR4
Streptococcus pyogenes MGAS10394
Streptococcus pyogenes MGAS315
Streptococcus pyogenes MGAS8232
Streptococcus pyogenes SF370
Streptococcus pyogenes SSI-1
Streptococcus thermophilus ATCC BAA-250
Streptococcus thermophilus CNRZ 1066
Streptomyces avermitilis
Sulfolobus tokodaii
Symbiobacterium thermophilum
Synechococcus elongatus
Synechococcus sp. PCC 6301
Synechococcus sp. WH8102
Synechocystis sp.
Thermoanaerobacter tengcongensis
Thermoplasma volcanium
Thermotoga maritima
Thermus thermophilus HB27
Treponema denticola
Treponema pallidum
Tropheryma whipplei Twist
Ureaplasma parvum
Vibrio vulnificus YJ016
Wigglesworthia glossinidia brevipalpis
Wolbachia sp.
Wolinella succinogenes
Xanthomonas axonopodis
Xylella fastidiosa Temecula1
Yersinia pseudotuberculosis
Zymomonas mobilis


Archaea
Nanoarchaeota
Crenarchaeota
Euryarchaeota
Nanoarchaeum
Thermoprotei
Methanomicrobia
Methanobacteria
Archaeoglobi
Methanococci
Halobacteria
Methanopyri
Thermococci
Thermoplasmata
Thermotogae
Chlorobi
Planctomycetes
Thermotogales
Chlorobia
Epsilonbacterales
Planctomycetacia
Spirochaetes
Spirochaetales
Gloeobacteria
Nostocales
Cyanobacteria
Chroococcales
Prochlorales
Bacteroidetes
Bacteroides (class)
Actinobacteria
Actinobacteridae
Deinococcus–Thermus
Symbiobacterium
Thermacocci
Aquificae
Aquificales
Chlamydiae
Chlamydiales
Bacteria
Epsilonproteobacteria
Deltaproteobacteria
Alphaproteobacteria
Betaproteobacteria
Proteobacteria
Gammaproteobacteria
Clostridia
Mollicutes
Bacillales
Firmicutes
Lactobacillales

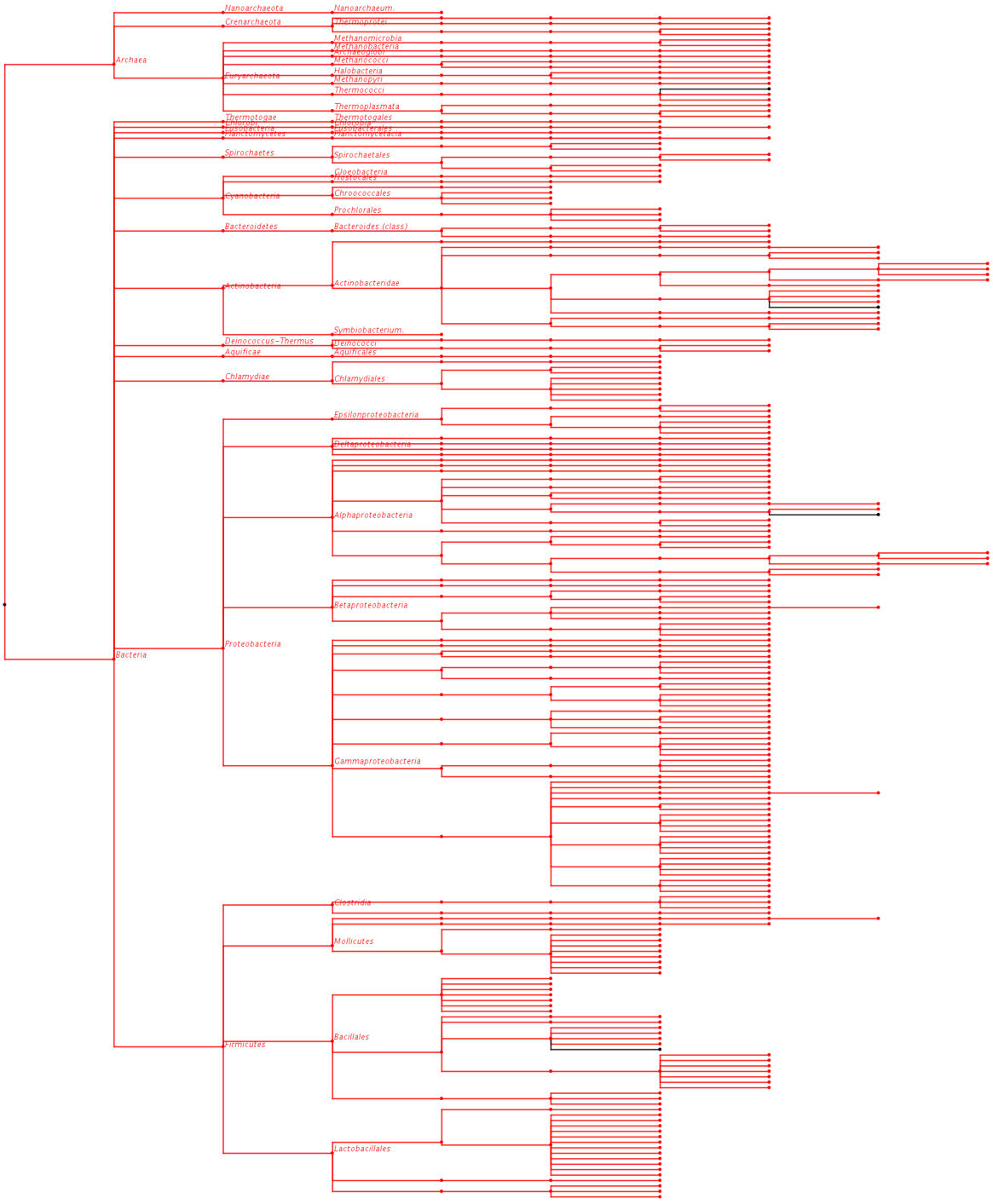
Level 7

Acinetobacter sp.
Aeropyrum pernix
Anabaena sp.
Anaplasma marginale
Aquifex aeolicus
Archaeoglobus fulgidus
Azoarcus sp.
Bacillus anthracis 0581
Bacillus anthracis Porton
Bacillus anthracis Sterne
Bacillus cereus ATCC 10987
Bacillus cereus ATCC 14579
Bacillus cereus ZK
Bacillus clausii
Bacillus halodurans
Bacillus licheniformis Goettingen
Bacillus subtilis
Bacillus thuringiensis
Bacteroides fragilis YCH46
Bacteroides thetaiotaomicron
Bartonella henselae
Bartonella quintana
Bdellovibrio bacteriovorus
Bifidobacterium longum
Bordetella bronchiseptica
Bordetella parapertussis
Bordetella pertussis
Borrelia burgdorferi
Borrelia garinii
Bradyrhizobium japonicum
Brucella melitensis
Brucella suis
Buchnera aphidicola Acyrthosiphon pisum
Buchnera aphidicola Baizongia pistaciae
Buchnera aphidicola Schizaphis graminum
Burkholderia mallei
Burkholderia pseudomallei
Campylobacter jejuni NCTC 11168
Campylobacter jejuni RM1221
Candidatus Blochmannia floridanus
Caulobacter crescentus
Chlamydia muridarum
Chlamydia pneumoniae AR39
Chlamydia pneumoniae CWL029
Chlamydia pneumoniae J138
Chlamydia pneumoniae TW-183
Chlamydia trachomatis
Chlamydophila caviae
Chlorobium tepidum
Chromobacterium violaceum
Clostridium acetobutylicum
Clostridium perfringens
Clostridium tetani
Corynebacterium glutamicum Nakagawa
Coxiella burnetii
Deinococcus radiodurans
Desulfotalea psychrophila
Desulfovibrio vulgaris
Ehrlichia ruminantium CIRAD
Ehrlichia ruminantium Gardel
Enterococcus faecalis
Erwinia carotovora
Escherichia coli EDL933
Escherichia coli K12
Escherichia coli O6 UPEC
Escherichia coli Sakai
Francisella tularensis
Fusobacterium nucleatum
Geobacillus kaustophilus
Geobacter sulfurreducens
Gloeobacter violaceus
Gluconobacter oxydans
Haemophilus ducreyi
Haemophilus influenzae ATCC 51907
Haloarcula marismortui
Halobacterium salinarium
Helicobacter hepaticus
Helicobacter pylori ATCC 700392
Helicobacter pylori J99
Idiomarina loihiensis
Lactobacillus acidophilus
Lactobacillus johnsonii
Lactobacillus plantarum
Lactococcus lactis
Legionella pneumophila Lens
Legionella pneumophila Paris
Legionella pneumophila Philadelphia 1
Leifsonia xyli
Leptospira interrogans Icterohaemorrhagiae
Leptospira interrogans lai
Listeria innocua
Listeria monocytogenes 1/2a
Listeria monocytogenes 4b
Mannheimia succiniciproducens
Mesoplasma florum
Methanobacterium thermoautotrophicum
Methanococcus jannaschii
Methanococcus maripaludis
Methanopyrus kandleri
Methanosarcina acetivorans
Methanosarcina mazei
Methylococcus capsulatus
Mycobacterium paratuberculosis
Mycoplasma gallisepticum
Mycoplasma genitalium
Mycoplasma hyopneumoniae
Mycoplasma mobile
Mycoplasma mycoides
Mycoplasma penetrans
Mycoplasma pneumoniae
Mycoplasma pulmonis
Nanoarchaeum equitans
Neisseria meningitidis A
Neisseria meningitidis B
Nitrosomonas europaea
Nocardia farcinica
Oceanobacillus iheyensis
Onion yellows phytoplasma
Parachlamydia sp.
Pasteurella multocida
Photobacterium profundum
Photorhabdus luminescens
Picrophilus torridus
Porphyromonas gingivalis
Prochlorococcus marinus CCMP 1375
Prochlorococcus marinus CCMP 1378
Prochlorococcus marinus MIT 9313
Propionibacterium acnes
Pseudomonas aeruginosa
Pseudomonas putida
Pseudomonas syringae tomato
Pyrobaculum aerophilum
Pyrococcus furiosus
Pyrococcus horikoshii
Ralstonia solanacearum
Rhizobium loti
Rhizobium meliloti
Rhodopirellula baltica
Rhodopseudomonas palustris
Rickettsia conorii
Salmonella paratyphi-a
Salmonella typhi ATCC 700931
Salmonella typhi CT18
Salmonella typhimurium
Shewanella oneidensis
Shigella flexneri 2457T
Shigella flexneri 301
Silicibacter pomeroyi
Staphylococcus aureus COL
Staphylococcus aureus MRSA252
Staphylococcus aureus MSSA476
Staphylococcus aureus MW2
Staphylococcus aureus Mu50
Staphylococcus aureus N315
Staphylococcus epidermidis ATCC 12228
Streptococcus agalactiae III
Streptococcus agalactiae V
Streptococcus mutans
Streptococcus pneumoniae ATCC BAA-255
Streptococcus pneumoniae TIGR4
Streptococcus pyogenes MGAS10394
Streptococcus pyogenes MGAS315
Streptococcus pyogenes MGAS8232
Streptococcus pyogenes SF370
Streptococcus pyogenes SSI-1
Streptococcus thermophilus ATCC BAA-250
Streptococcus thermophilus CNRZ 1066
Streptomyces avermitilis
Sulfolobus solfataricus
Sulfolobus tokodaii
Symbiobacterium thermophilum
Synechococcus elongatus
Synechococcus sp. PCC 6301
Synechococcus sp. WH8102
Synechocystis sp.
Thermoanaerobacter tengcongensis
Thermoplasma acidophilum
Thermoplasma volcanium
Thermotoga maritima
Thermus thermophilus HB27
Thermus thermophilus HB8
Treponema denticola
Treponema pallidum
Tropheryma whipplei Twist
Ureaplasma parvum
Vibrio cholerae
Vibrio parahaemolyticus
Vibrio vulnificus CMCP6
Vibrio vulnificus YJ016
Wigglesworthia glossinidia brevipalpis
Wolbachia sp.
Wolinella succinogenes
Xanthomonas axonopodis
Xanthomonas campestris campestris
Xanthomonas oryzae
Xylella fastidiosa 9a5c
Xylella fastidiosa Temecula1
Yersinia pestis 91001
Yersinia pestis CO-92
Yersinia pestis KIM5
Yersinia pseudotuberculosis
Zymomonas mobilis


Archaea
Nanoarchaeota
Crenarchaeota
Euryarchaeota
Nanoarchaeum
Thermoprotei
Methanomicrobia
Methanobacteria
Archaeoglobi
Methanococci
Halobacteria
Methanopyri
Thermococci
Thermoplasmata
Thermotogae
Chlorobi
Planctomycetes
Thermotogales
Chlorobia
Spirochaetales
Gloeobacteria
Nostocales
Chroococcales
Prochlorales
Spirochaetes
Cyanobacteria
Bacteroidetes
Bacteroides (class)
Actinobacteria
Actinobacteridae
Deinococcus–Thermus
Aquificae
Chlamydiae
Symbiobacterium
Thermacoci
Aquificales
Chlamydiales
Bacteria
Proteobacteria
Epsilonproteobacteria
Deltaproteobacteria
Alphaproteobacteria
Betaproteobacteria
Gammaproteobacteria
Firmicutes
Clostridia
Mollicutes
Bacillales
Lactobacillales

Level 8

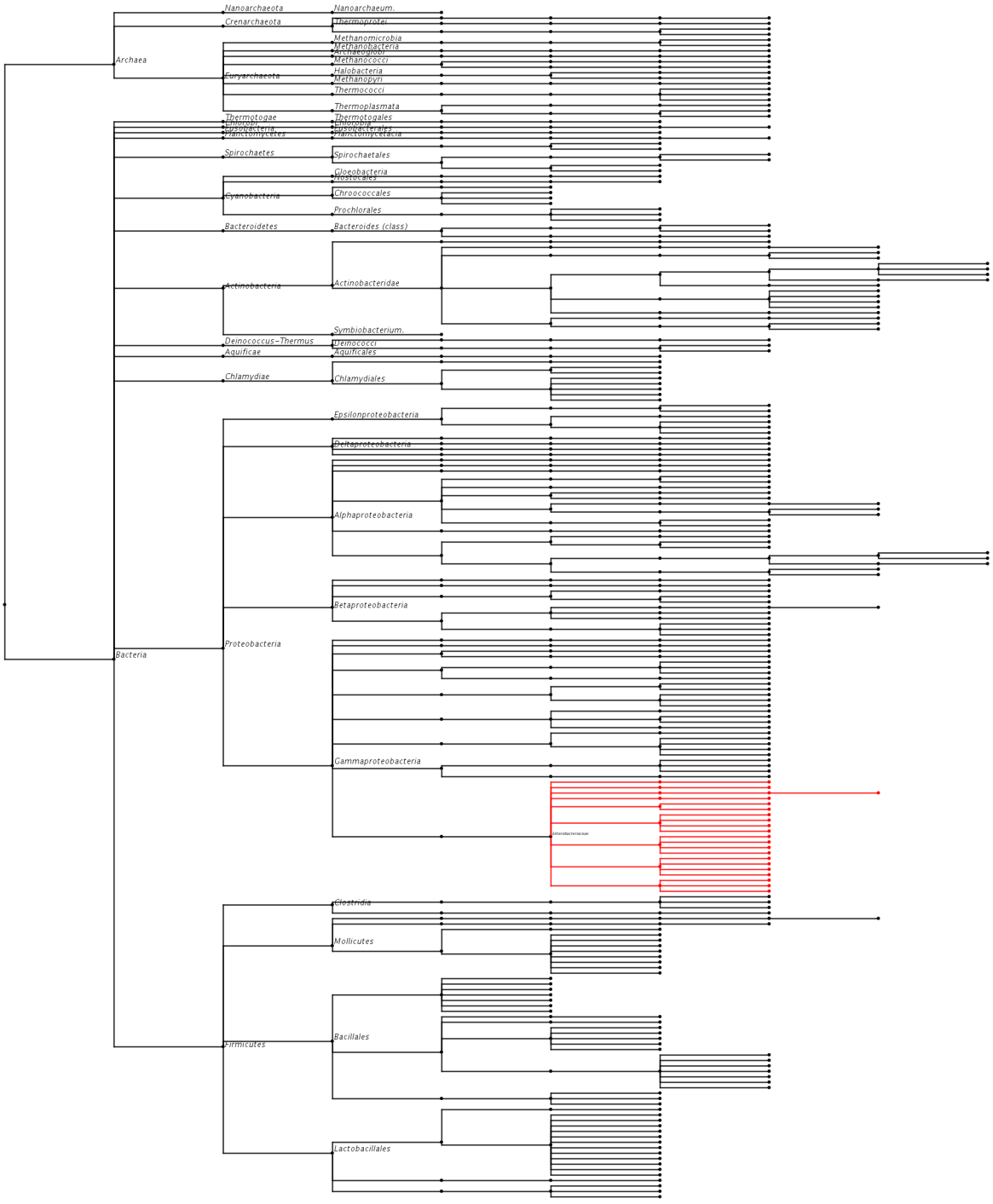
Acinetobacter sp.
Aeropyrum pernix
Agrobacterium tumefaciens Dupont
Anabaena sp.
Anaplasma marginale
Aquifex aeolicus
Archaeoglobus fulgidus
Azoarcus sp.
Bacillus anthracis 0581
Bacillus anthracis Porton
Bacillus anthracis Sterne
Bacillus cereus ATCC 10987
Bacillus cereus ATCC 14579
Bacillus cereus ZK
Bacillus clausii
Bacillus halodurans
Bacillus licheniformis Goettingen
Bacillus subtilis
Bacillus thuringiensis
Bacteroides fragilis YCH46
Bacteroides thetaiotaomicron
Bartonella henselae
Bartonella quintana
Bdellovibrio bacteriovorus
Bifidobacterium longum
Bordetella bronchiseptica
Bordetella parapertussis
Bordetella pertussis
Borrelia burgdorferi
Borrelia garinii
Bradyrhizobium japonicum
Brucella melitensis
Brucella suis
Buchnera aphidicola Acyrthosiphon pisum
Buchnera aphidicola Baizongia pistaciae
Buchnera aphidicola Schizaphis graminum
Burkholderia mallei
Burkholderia pseudomallei
Campylobacter jejuni NCTC 11168
Campylobacter jejuni RM1221
Candidatus Blochmannia floridanus
Caulobacter crescentus
Chlamydia muridarum
Chlamydia pneumoniae AR39
Chlamydia pneumoniae CWL029
Chlamydia pneumoniae J138
Chlamydia pneumoniae TW-183
Chlamydia trachomatis
Chlamydophila caviae
Chlorobium tepidum
Chromobacterium violaceum
Clostridium acetobutylicum
Clostridium perfringens
Clostridium tetani
Corynebacterium diphtheriae
Corynebacterium efficiens
Corynebacterium glutamicum Nakagawa
Coxiella burnetii
Deinococcus radiodurans
Desulfotalea psychrophila
Desulfovibrio vulgaris
Ehrlichia ruminantium CIRAD
Ehrlichia ruminantium Gardel
Enterococcus faecalis
Erwinia carotovora
Escherichia coli EDL933
Escherichia coli K12
Escherichia coli O6 UPEC
Escherichia coli Sakai
Francisella tularensis
Fusobacterium nucleatum
Geobacillus kaustophilus
Geobacter sulfurreducens
Gloeobacter violaceus
Gluconobacter oxydans
Haemophilus ducreyi
Haemophilus influenzae ATCC 51907
Haloarcula marismortui
Halobacterium salinarium
Helicobacter hepaticus
Helicobacter pylori ATCC 700392
Helicobacter pylori J99
Idiomarina loihiensis
Lactobacillus acidophilus
Lactobacillus johnsonii
Lactobacillus plantarum
Lactococcus lactis
Legionella pneumophila Lens
Legionella pneumophila Paris
Legionella pneumophila Philadelphia 1
Leifsonia xyli
Leptospira interrogans Icterohaemorrhagiae
Leptospira interrogans lai
Listeria innocua
Listeria monocytogenes 1/2a
Listeria monocytogenes 4b
Mannheimia succiniciproducens
Mesoplasma florum
Methanobacterium thermoautotrophicum
Methanococcus jannaschii
Methanococcus maripaludis
Methanopyrus kandleri
Methanosarcina acetivorans
Methanosarcina mazei
Methylococcus capsulatus
Mycobacterium leprae
Mycobacterium paratuberculosis
Mycobacterium tuberculosis Oshkosh
Mycoplasma gallisepticum
Mycoplasma genitalium
Mycoplasma hyopneumoniae
Mycoplasma mobile
Mycoplasma mycoides
Mycoplasma penetrans
Mycoplasma pneumoniae
Mycoplasma pulmonis
Nanoarchaeum equitans
Neisseria meningitidis A
Neisseria meningitidis B
Nitrosomonas europaea
Nocardia farcinica
Oceanobacillus iheyensis
Onion yellows phytoplasma
Parachlamydia sp.
Pasteurella multocida
Photobacterium profundum
Photorhabdus luminescens
Picrophilus torridus
Porphyromonas gingivalis
Prochlorococcus marinus CCMP 1375
Prochlorococcus marinus CCMP 1378
Prochlorococcus marinus MIT 9313
Propionibacterium acnes
Pseudomonas aeruginosa
Pseudomonas putida
Pseudomonas syringae tomato
Pyrobaculum aerophilum
Pyrococcus furiosus
Pyrococcus horikoshii
Ralstonia solanacearum
Rhizobium loti
Rhizobium meliloti
Rhodopirellula baltica
Rhodopseudomonas palustris
Rickettsia conorii
Rickettsia prowazekii
Salmonella paratyphi-a
Salmonella typhi ATCC 700931
Salmonella typhi CT18
Salmonella typhimurium
Shewanella oneidensis
Shigella flexneri 2457T
Shigella flexneri 301
Silicibacter pomeroyi
Staphylococcus aureus COL
Staphylococcus aureus MRSA252
Staphylococcus aureus MSSA476
Staphylococcus aureus MW2
Staphylococcus aureus Mu50
Staphylococcus aureus N315
Staphylococcus epidermidis ATCC 12228
Streptococcus agalactiae III
Streptococcus agalactiae V
Streptococcus mutans
Streptococcus pneumoniae ATCC BAA-255
Streptococcus pneumoniae TIGR4
Streptococcus pyogenes MGAS10394
Streptococcus pyogenes MGAS315
Streptococcus pyogenes MGAS8232
Streptococcus pyogenes SF370
Streptococcus pyogenes SSI-1
Streptococcus thermophilus ATCC BAA-250
Streptococcus thermophilus CNRZ 1066
Streptomyces avermitilis
Streptomyces coelicolor
Sulfolobus solfataricus
Sulfolobus tokodaii
Symbiobacterium thermophilum
Synechococcus elongatus
Synechococcus sp. PCC 6301
Synechococcus sp. WH8102
Synechocystis sp.
Thermoanaerobacter tengcongensis
Thermoplasma acidophilum
Thermoplasma volcanium
Thermotoga maritima
Thermus thermophilus HB27
Thermus thermophilus HB8
Treponema denticola
Treponema pallidum
Tropheryma whipplei TW08/27
Tropheryma whipplei Twist
Ureaplasma parvum
Vibrio cholerae
Vibrio parahaemolyticus
Vibrio vulnificus CMCP6
Vibrio vulnificus YJ016
Wigglesworthia glossinidia brevipalpis
Wolbachia pipientis wMel
Wolbachia sp.
Wolinella succinogenes
Xanthomonas axonopodis
Xanthomonas campestris campestris
Xanthomonas oryzae
Xylella fastidiosa 9a5c
Xylella fastidiosa Temecula1
Yersinia pestis 91001
Yersinia pestis CO-92
Yersinia pestis KIM5
Yersinia pseudotuberculosis
Zymomonas mobilis



Level 9 = Nearest 7 = All

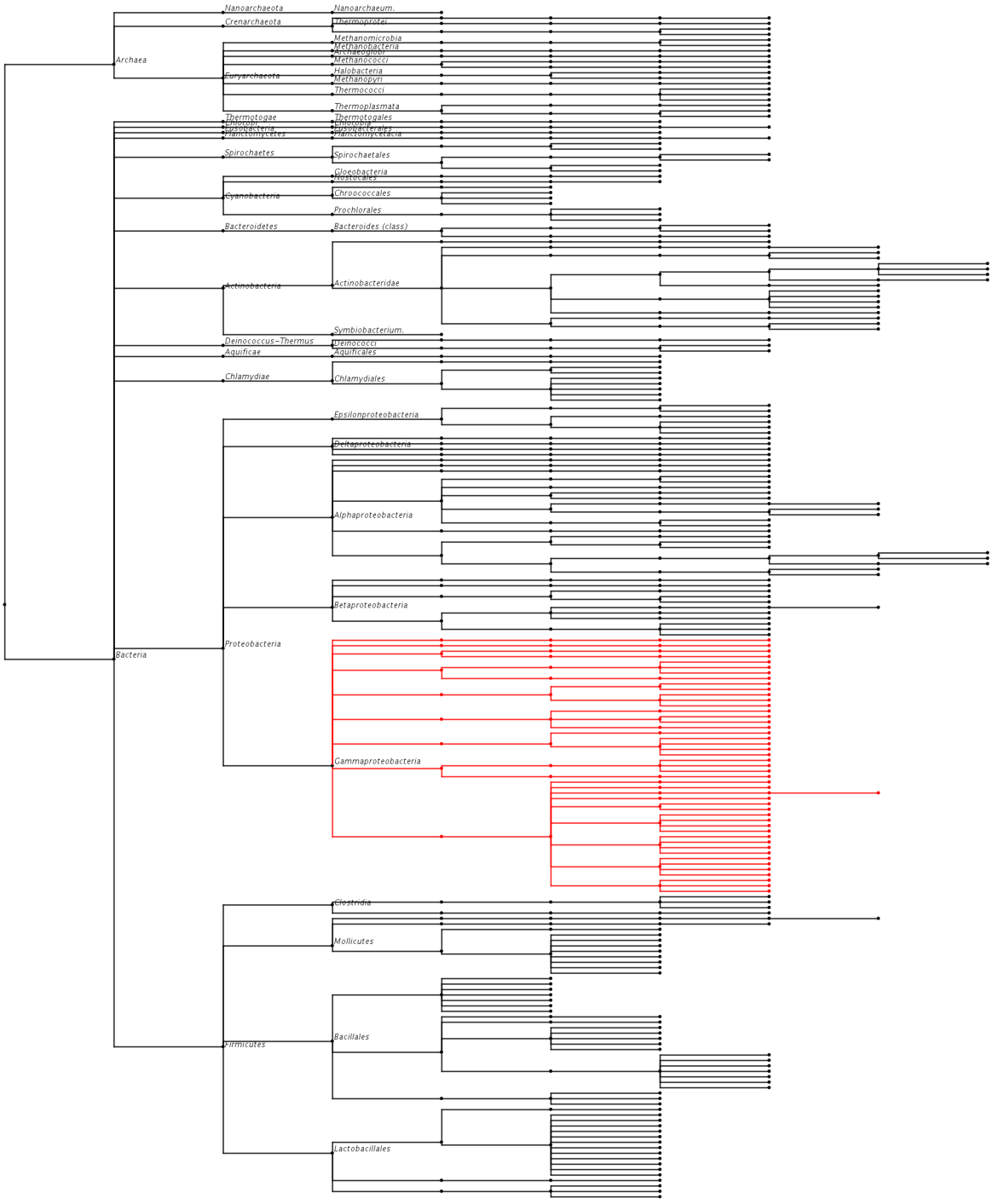
Acinetobacter sp.
Aeropyrum pernix
Agrobacterium tumefaciens
Anabaena sp.
Anaplasma marginale
Aquifex aeolicus
Archaeoglobus fulgidus
Azoarcus sp.
Bacillus anthracis
Bacillus anthracis
Bacillus anthracis
Bacillus cereus
Bacillus cereus
Bacillus cereus
Bacillus clausii
Bacillus halodurans
Bacillus licheniformis
Bacillus subtilis
Bacillus thuringiensis
Bacteroides fragilis
Bacteroides thetaiotaomicron
Bartonella henselae
Bartonella quintana
Bdellovibrio bacteriovorus
Bifidobacterium longum
Bordetella bronchiseptica
Bordetella parapertussis
Bordetella pertussis
Borrelia burgdorferi
Borrelia garinii
Bradyrhizobium japonicum
Brucella melitensis
Brucella suis
Buchnera aphidicola
Buchnera aphidicola
Buchnera aphidicola
Burkholderia mallei
Burkholderia pseudomallei
Campylobacter jejuni
Campylobacter jejuni
Candidatus Blochmannia
Caulobacter crescentus
Chlamydia muridarum
Chlamydia pneumoniae
Chlamydia pneumoniae
Chlamydia pneumoniae
Chlamydia pneumoniae
Chlamydia trachomatis
Chlamydophila caviae
Chlorobium tepidum
Chromobacterium violaceum
Clostridium acetobutylicum
Clostridium perfringens
Clostridium tetani
Corynebacterium diphtheriae
Corynebacterium efficiens
Corynebacterium glutamicum
Coxiella burnetii
Deinococcus radiodurans
Desulfotalea psychrophila
Desulfovibrio vulgaris
Ehrlichia ruminantium
Ehrlichia ruminantium
Enterococcus faecalis
Erwinia carotovora
Escherichia coli
Escherichia coli
Escherichia coli
Escherichia coli
Francisella tularensis
Fusobacterium nucleatum
Geobacillus kaustophilus
Geobacter sulfurreducens
Gloeobacter violaceus
Gluconobacter oxydans
Haemophilus ducreyi
Haemophilus influenzae
Haloarcula marismortui
Halobacterium salinarium
Helicobacter hepaticus
Helicobacter pylori
Helicobacter pylori
Idiomarina loihiensis
Lactobacillus acidophilus
Lactobacillus johnsonii
Lactobacillus plantarum
Lactococcus lactis
Legionella pneumophila
Legionella pneumophila
Legionella pneumophila
Leifsonia xyli
Leptospira interrogans
Leptospira interrogans
Listeria innocua
Listeria monocytogenes
Listeria monocytogenes
Mannheimia succiniciproducens
Mesoplasma florum
Methanobacterium thermoautotrophicum
Methanococcus jannaschii
Methanococcus maripaludis
Methanopyrus kandleri
Methanosarcina acetivorans
Methanosarcina mazei
Methylococcus capsulatus
Mycobacterium bovis
Mycobacterium leprae
Mycobacterium paratuberculosis
Mycobacterium tuberculosis
Mycobacterium tuberculosis
Mycoplasma gallisepticum
Mycoplasma genitalium
Mycoplasma hyopneumoniae
Mycoplasma mobile
Mycoplasma mycoides
Mycoplasma penetrans
Mycoplasma pneumoniae
Mycoplasma pulmonis
Nanoarchaeum equitans
Neisseria meningitidis
Neisseria meningitidis
Nitrosomonas europaea
Nocardia farcinica
Oceanobacillus iheyensis
Onion yellows
Parachlamydia sp.
Pasteurella multocida
Photobacterium profundum
Photorhabdus luminescens
Picrophilus torridus
Porphyromonas gingivalis
Prochlorococcus marinus
Prochlorococcus marinus
Prochlorococcus marinus
Propionibacterium acnes
Pseudomonas aeruginosa
Pseudomonas putida
Pseudomonas syringae
Pyrobaculum aerophilum
Pyrococcus furiosus
Pyrococcus horikoshii
Ralstonia solanacearum
Rhizobium loti
Rhizobium meliloti
Rhodopirellula baltica
Rhodopseudomonas palustris
Rickettsia conorii
Rickettsia prowazekii
Rickettsia typhi
Salmonella paratyphi-a
Salmonella typhi
Salmonella typhi
Salmonella typhimurium
Shewanella oneidensis
Shigella flexneri
Shigella flexneri
Silicibacter pomeroyi
Staphylococcus aureus
Staphylococcus aureus
Staphylococcus aureus
Staphylococcus aureus
Staphylococcus aureus
Staphylococcus aureus
Staphylococcus epidermidis
Streptococcus agalactiae
Streptococcus agalactiae
Streptococcus mutans
Streptococcus pneumoniae
Streptococcus pneumoniae
Streptococcus pyogenes
Streptococcus pyogenes
Streptococcus pyogenes
Streptococcus pyogenes
Streptococcus pyogenes
Streptococcus thermophilus
Streptococcus thermophilus
Streptomyces avermitilis
Streptomyces coelicolor
Sulfolobus solfataricus
Sulfolobus tokodaii
Symbiobacterium thermophilum
Synechococcus elongatus
Synechococcus sp.
Synechococcus sp.
Synechocystis sp.
Thermoanaerobacter tengcongensis
Thermoplasma acidophilum
Thermoplasma volcanium
Thermotoga maritima
Thermus thermophilus
Thermus thermophilus
Treponema denticola
Treponema pallidum
Tropheryma whipplei
Tropheryma whipplei
Ureaplasma parvum
Vibrio cholerae
Vibrio parahaemolyticus
Vibrio vulnificus
Vibrio vulnificus
Wigglesworthia glossinidia
Wolbachia pipientis
Wolbachia sp.
Wolinella succinogenes
Xanthomonas axonopodis
Xanthomonas campestris
Xanthomonas oryzae
Xylella fastidiosa
Xylella fastidiosa
Yersinia pestis
Yersinia pestis
Yersinia pestis
Yersinia pseudotuberculosis
Zymomonas mobilis

Level 9 = Nearest 7 = All


Archaea
Nanoarchaeota
Crenarchaeota
Euryarchaeota
Thermotogae
Chlorobi
Planctomycetes
Spirochaetes
Cyanobacteria
Bacteroidetes
Actinobacteria
Deinococcus–Thermus
Aquificae
Chlamydiae
Bacteria
Proteobacteria
Firmicutes
Nanoarchaeum
Thermoprotei
Methanomicrobia
Methanobacteria
Archaeoglobi
Methanococci
Halobacteria
Methanopyri
Thermococci
Thermoplasmata
Thermotogales
Chlorobia
Sphaerobacterales
Planctomycetacia
Spirochaetales
Gloeobacteria
Nostocales
Chroococcales
Prochlorales
Bacteroides (class)
Actinobacteridae
Symbiobacterium
Deinococci
Aquificales
Chlamydiales
Epsilonproteobacteria
Deltaproteobacteria
Alphaproteobacteria
Betaproteobacteria
Gammaproteobacteria
Enterobacteriaceae
Clostridia
Mollicutes
Bacillales
Lactobacillales

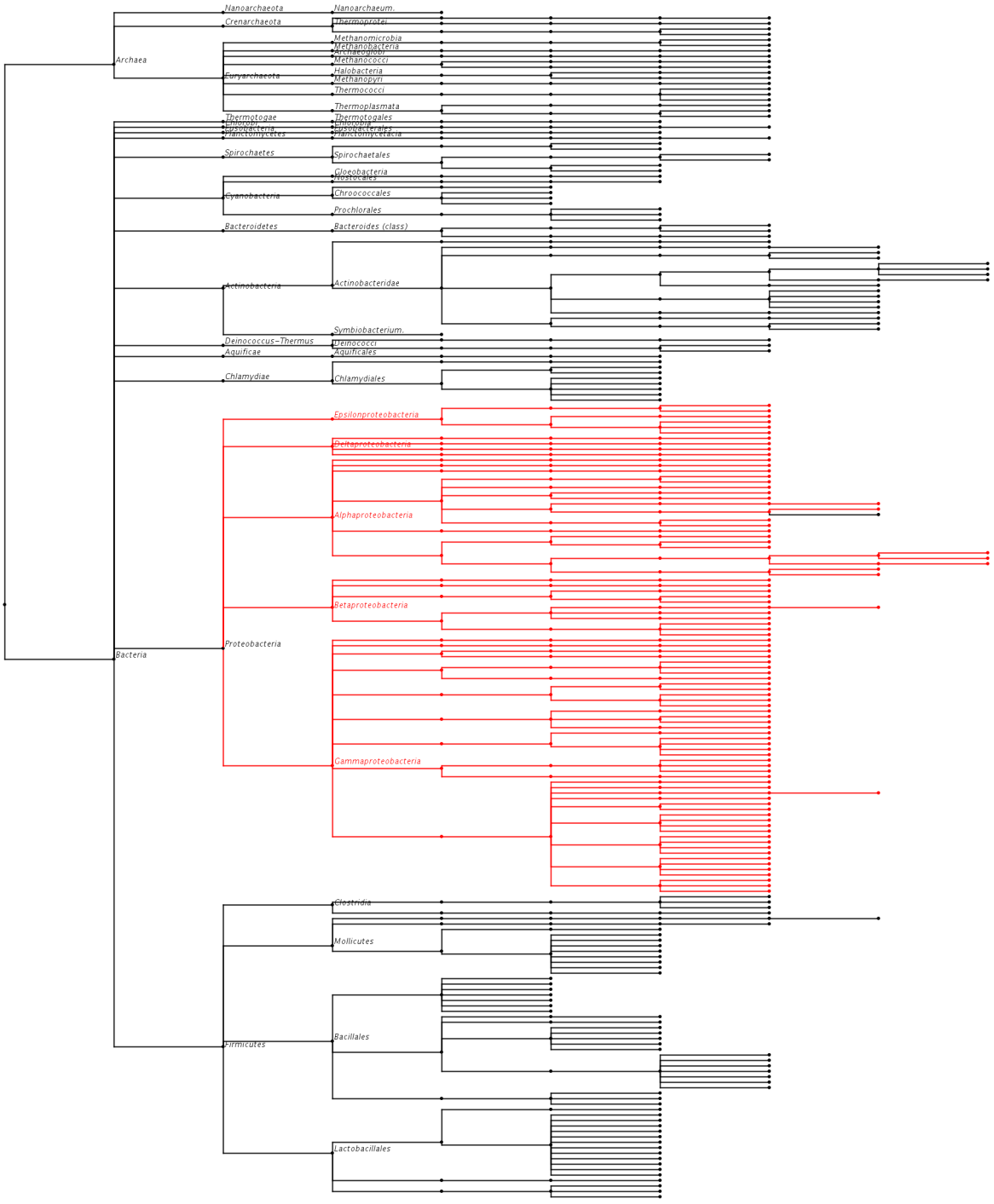
Nearest 2

Buchnera aphidicola Acyrthosiphon pisum
Buchnera aphidicola Baizongia pistaciae
Buchnera aphidicola Schizaphis graminum
Candidatus Blochmannia floridanus
Erwinia carotovora
Escherichia coli EDL933
Escherichia coli K12
Escherichia coli O6 UPEC
Escherichia coli Sakai
Photorhabdus luminescens
Salmonella paratyphi-a
Salmonella typhi ATCC 700931
Salmonella typhi CT18
Salmonella typhimurium
Shigella flexneri 2457T
Shigella flexneri 301
Wigglesworthia glossinidia brevipalpis
Yersinia pestis 91001
Yersinia pestis CO-92
Yersinia pestis KIM5
Yersinia pseudotuberculosis



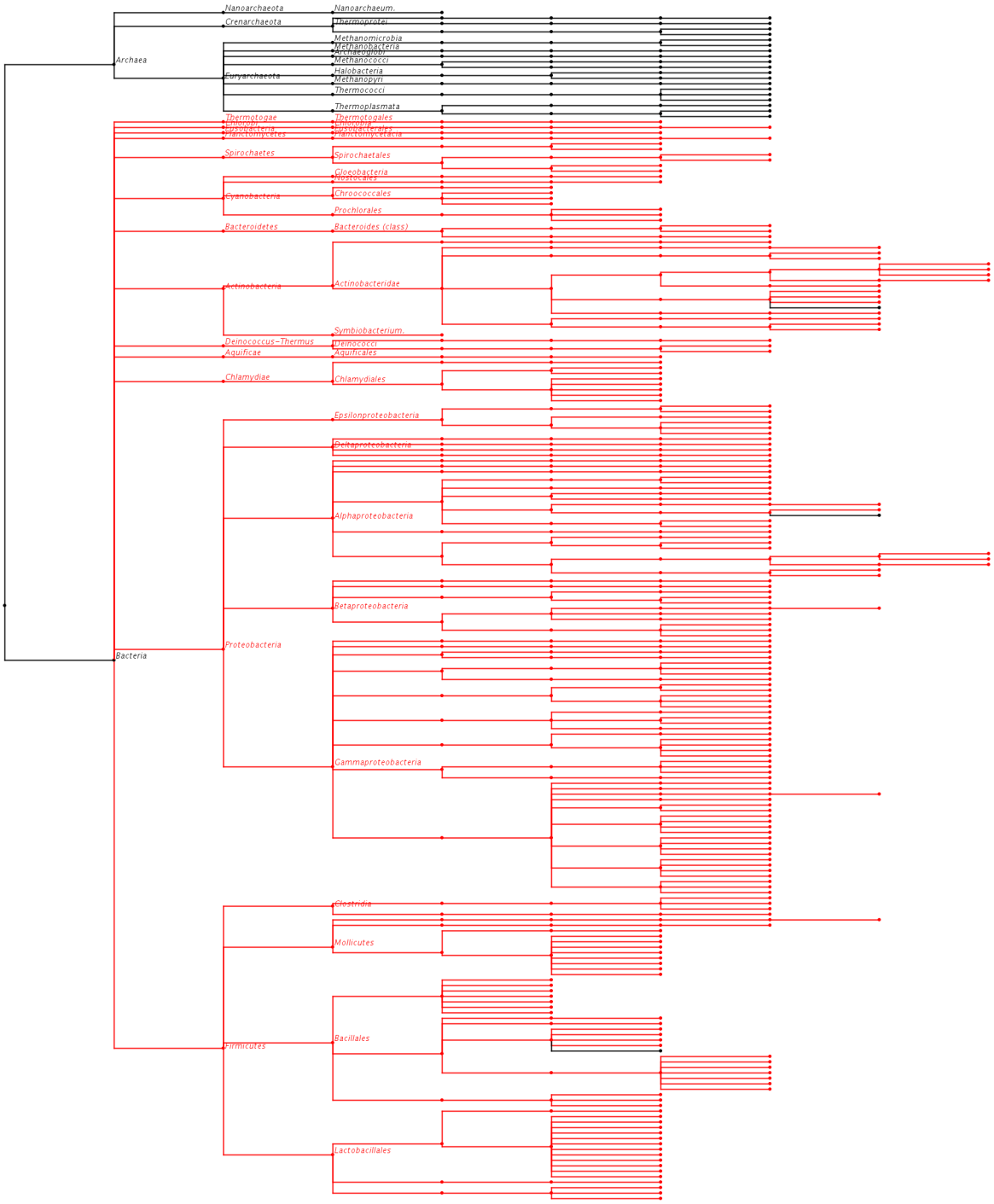
Nearest 3 = Nearest 4

Acinetobacter sp.
Buchnera aphidicola Acyrthosiphon pisum
Buchnera aphidicola Baizongia pistaciae
Buchnera aphidicola Schizaphis graminum
Candidatus Blochmannia floridanus
Coxiella burnetii
Erwinia carotovora
Escherichia coli EDL933
Escherichia coli K12
Escherichia coli O6 UPEC
Escherichia coli Sakai
Francisella tularensis
Haemophilus ducreyi
Haemophilus influenzae ATCC 51907
Idiomarina loihiensis
Legionella pneumophila Lens
Legionella pneumophila Paris
Legionella pneumophila Philadelphia 1
Mannheimia succiniciproducens
Methylococcus capsulatus
Pasteurella multocida
Photobacterium profundum
Photorhabdus luminescens
Pseudomonas aeruginosa
Pseudomonas putida
Pseudomonas syringae tomato
Salmonella paratyphi-a
Salmonella typhi ATCC 700931
Salmonella typhi CT18
Salmonella typhimurium
Shewanella oneidensis
Shigella flexneri 2457T
Shigella flexneri 301
Vibrio cholerae
Vibrio parahaemolyticus
Vibrio vulnificus CMCP6
Vibrio vulnificus YJ016
Wigglesworthia glossinidia brevipalpis
Xanthomonas axonopodis
Xanthomonas campestris campestris
Xanthomonas oryzae
Xylella fastidiosa 9a5c
Xylella fastidiosa Temecula1
Yersinia pestis 91001
Yersinia pestis CO-92
Yersinia pestis KIM5
Yersinia pseudotuberculosis


Archaea
Nanoarchaeota
Crenarchaeota
Euryarchaeota
Nanoarchaeum
Thermoprotei
Methanomicrobia
Methanobacteria
Archaeoglobi
Methanococci
Halobacteria
Methanopyri
Thermococci
Thermoplasmata
Thermotogae
Chlorobi
Planctomycetes
Thermotogales
Chlorobia
Sphaerobacterales
Planctomycetacia
Spirochaetes
Spirochaetales
Cyanobacteria
Gloeobacteria
Nostocales
Chroococcales
Prochlorales
Bacteroidetes
Bacteroides (class)
Actinobacteria
Actinobacteridae
Deinococcus–Thermus
Aquificae
Symbiobacterium
Deinococci
Aquificales
Chlamydiae
Chlamydiales
Bacteria
Proteobacteria
Epsilonproteobacteria
Deltaproteobacteria
Alphaproteobacteria
Betaproteobacteria
Gammaproteobacteria
Firmicutes
Clostridia
Mollicutes
Bacillales
Lactobacillales

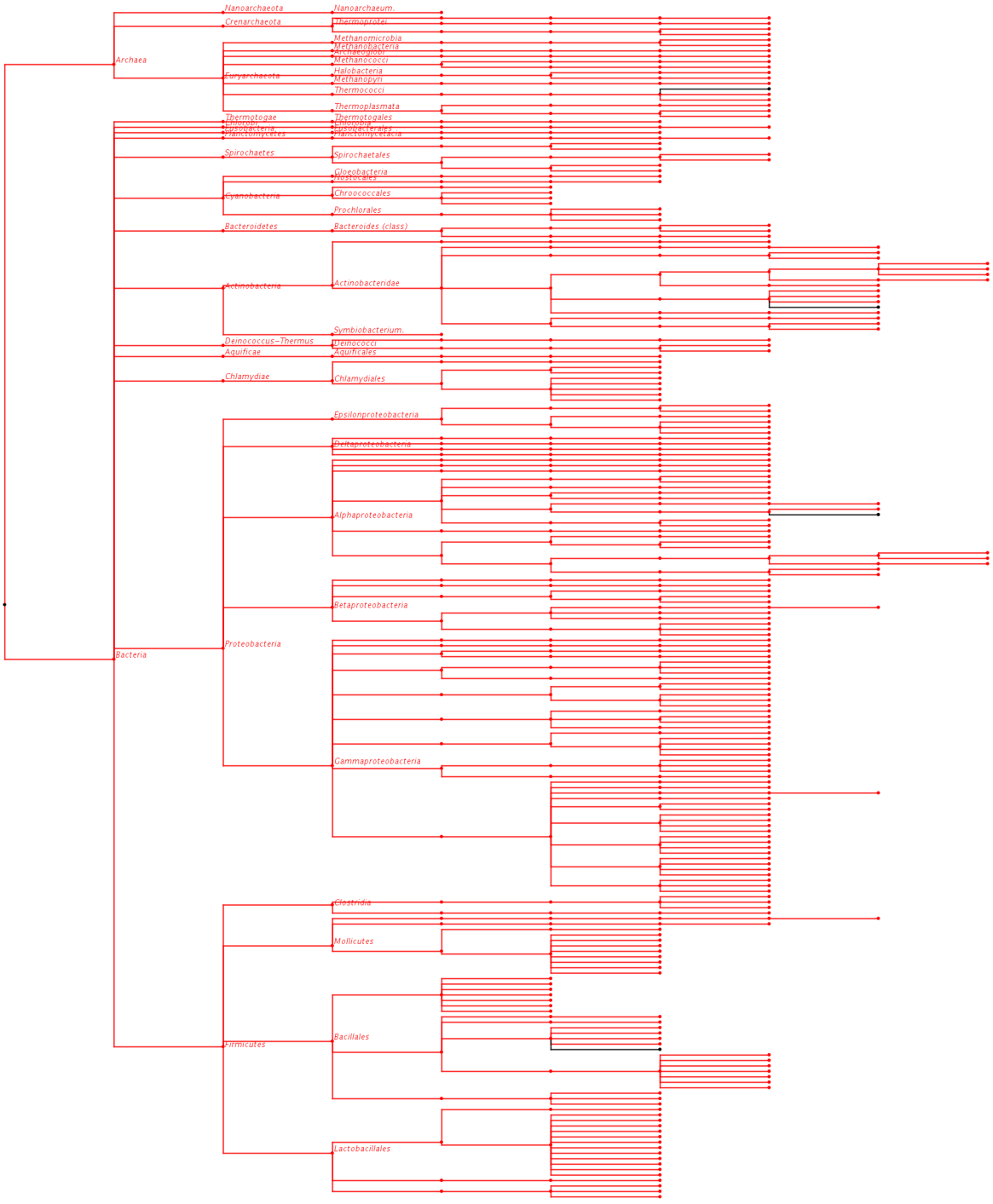
Nearest 5

Acinetobacter sp.
Agrobacterium tumefaciens Dupont
Anaplasma marginale
Azoarcus sp.
Bartonella henselae
Bartonella quintana
Bdellovibrio bacteriovorus
Bordetella bronchiseptica
Bordetella parapertussis
Bordetella pertussis
Bradyrhizobium japonicum
Brucella melitensis
Brucella suis
Buchnera aphidicola Acyrthosiphon pisum
Buchnera aphidicola Baizongia pistaciae
Buchnera aphidicola Schizaphis graminum
Burkholderia mallei
Burkholderia pseudomallei
Campylobacter jejuni NCTC 11168
Campylobacter jejuni RM1221
Candidatus Blochmannia floridanus
Caulobacter crescentus
Chromobacterium violaceum
Coxiella burnetii
Desulfotalea psychrophila
Desulfovibrio vulgaris
Ehrlichia ruminantium CIRAD
Ehrlichia ruminantium Gardel
Erwinia carotovora
Escherichia coli EDL933
Escherichia coli K12
Escherichia coli O6 UPEC
Escherichia coli Sakai
Francisella tularensis
Geobacter sulfurreducens
Gluconobacter oxydans
Haemophilus ducreyi
Haemophilus influenzae ATCC 51907
Helicobacter hepaticus
Helicobacter pylori ATCC 700392
Helicobacter pylori J99
Idiomarina loihiensis
Legionella pneumophila Lens
Legionella pneumophila Paris
Legionella pneumophila Philadelphia 1
Mannheimia succiniciproducens
Methylococcus capsulatus
Neisseria meningitidis A
Neisseria meningitidis B
Nitrosomonas europaea
Pasteurella multocida
Photobacterium profundum
Photorhabdus luminescens
Pseudomonas aeruginosa
Pseudomonas putida
Pseudomonas syringae tomato
Ralstonia solanacearum
Rhizobium loti
Rhizobium meliloti
Rhodopseudomonas palustris
Rickettsia conorii
Rickettsia prowazekii
Rickettsia typhi
Salmonella paratyphi-a
Salmonella typhi ATCC 700931
Salmonella typhi CT18
Salmonella typhimurium
Shewanella oneidensis
Shigella flexneri 2457T
Shigella flexneri 301
Silicibacter pomeroyi
Vibrio cholerae
Vibrio parahaemolyticus
Vibrio vulnificus CMCP6
Vibrio vulnificus YJ016
Wigglesworthia glossinidia brevipalpis
Wolbachia pipientis wMel
Wolbachia sp.
Wolinella succinogenes
Xanthomonas axonopodis
Xanthomonas campestris campestris
Xanthomonas oryzae
Xylella fastidiosa 9a5c
Xylella fastidiosa Temecula1
Yersinia pestis 91001
Yersinia pestis CO-92
Yersinia pestis KIM5
Yersinia pseudotuberculosis
Zymomonas mobilis


Archaea
Nanoarchaeota
Crenarchaeota
Euryarchaeota
Nanoarchaeum
Thermoprotei
Methanomicrobia
Methanobacteria
Archaeoglobi
Methanococci
Halobacteria
Methanopyri
Thermococci
Thermoplasmata
Bacteria
Thermotogae
Chlorobi
Planctomycetes
Spirochaetes
Cyanobacteria
Bacteroidetes
Actinobacteria
Deinococcus–Thermus
Aquificae
Chlamydiae
Proteobacteria
Thermotogales
Chlorobia
Planctomycetacia
Spirochaetales
Gloeobacteria
Nostocales
Chroococcales
Prochlorales
Bacteroides (class)
Actinobacteridae
Symbiobacterium
Deinococci
Aquificales
Chlamydiales
Epsilonproteobacteria
Deltaproteobacteria
Alphaproteobacteria
Betaproteobacteria
Gammaproteobacteria
Clostridia
Mollicutes
Bacillales
Lactobacillales
Firmicutes

Nearest 6

Acinetobacter sp.
Agrobacterium tumefaciens Dupont
Anabaena sp.
Anaplasma marginale
Aquifex aeolicus
Azoarcus sp.
Bacillus anthracis 0581
Bacillus anthracis Porton
Bacillus anthracis Sterne
Bacillus cereus ATCC 10987
Bacillus cereus ATCC 14579
Bacillus cereus ZK
Bacillus clausii
Bacillus halodurans
Bacillus licheniformis Goettingen
Bacillus subtilis
Bacillus thuringiensis
Bacteroides fragilis YCH46
Bacteroides thetaiotaomicron
Bartonella henselae
Bartonella quintana
Bdellovibrio bacteriovorus
Bifidobacterium longum
Bordetella bronchiseptica
Bordetella parapertussis
Bordetella pertussis
Borrelia burgdorferi
Borrelia garinii
Bradyrhizobium japonicum
Brucella melitensis
Brucella suis
Buchnera aphidicola Acyrthosiphon pisum
Buchnera aphidicola Baizongia pistaciae
Buchnera aphidicola Schizaphis graminum
Burkholderia mallei
Burkholderia pseudomallei
Campylobacter jejuni NCTC 11168
Campylobacter jejuni RM1221
Candidatus Blochmannia floridanus
Caulobacter crescentus
Chlamydia muridarum
Chlamydia pneumoniae AR39
Chlamydia pneumoniae CWL029
Chlamydia pneumoniae J138
Chlamydia pneumoniae TW-183
Chlamydia trachomatis
Chlamydophila caviae
Chlorobium tepidum
Chromobacterium violaceum
Clostridium acetobutylicum
Clostridium perfringens
Clostridium tetani
Corynebacterium diphtheriae
Corynebacterium efficiens
Corynebacterium glutamicum Nakagawa
Coxiella burnetii
Deinococcus radiodurans
Desulfotalea psychrophila
Desulfovibrio vulgaris
Ehrlichia ruminantium CIRAD
Ehrlichia ruminantium Gardel
Enterococcus faecalis
Erwinia carotovora
Escherichia coli EDL933
Escherichia coli K12
Escherichia coli O6 UPEC
Escherichia coli Sakai
Francisella tularensis
Fusobacterium nucleatum
Geobacillus kaustophilus
Geobacter sulfurreducens
Gloeobacter violaceus
Gluconobacter oxydans
Haemophilus ducreyi
Haemophilus influenzae ATCC 51907
Helicobacter hepaticus
Helicobacter pylori ATCC 700392
Helicobacter pylori J99
Idiomarina loihiensis
Lactobacillus acidophilus
Lactobacillus johnsonii
Lactobacillus plantarum
Lactococcus lactis
Legionella pneumophila Lens
Legionella pneumophila Paris
Legionella pneumophila Philadelphia 1
Leifsonia xyli
Leptospira interrogans Icterohaemorrhagiae
Leptospira interrogans lai
Listeria innocua
Listeria monocytogenes 1/2a
Listeria monocytogenes 4b
Mannheimia succiniciproducens
Mesoplasma florum
Methylococcus capsulatus
Mycobacterium bovis
Mycobacterium leprae
Mycobacterium paratuberculosis
Mycobacterium tuberculosis H37Rv
Mycobacterium tuberculosis Oshkosh
Mycoplasma gallisepticum
Mycoplasma genitalium
Mycoplasma hyopneumoniae
Mycoplasma mobile
Mycoplasma mycoides
Mycoplasma penetrans
Mycoplasma pneumoniae
Mycoplasma pulmonis
Neisseria meningitidis A
Neisseria meningitidis B
Nitrosomonas europaea
Nocardia farcinica
Oceanobacillus iheyensis
Onion yellows phytoplasma
Parachlamydia sp.
Pasteurella multocida
Photobacterium profundum
Photorhabdus luminescens
Porphyromonas gingivalis
Prochlorococcus marinus CCMP 1375
Prochlorococcus marinus CCMP 1378
Prochlorococcus marinus MIT 9313
Propionibacterium acnes
Pseudomonas aeruginosa
Pseudomonas putida
Pseudomonas syringae tomato
Ralstonia solanacearum
Rhizobium loti
Rhizobium meliloti
Rhodopirellula baltica
Rhodopseudomonas palustris
Rickettsia conorii
Rickettsia prowazekii
Rickettsia typhi
Salmonella paratyphi-a
Salmonella typhi ATCC 700931
Salmonella typhi CT18
Salmonella typhimurium
Shewanella oneidensis
Shigella flexneri 2457T
Shigella flexneri 301
Silicibacter pomeroyi
Staphylococcus aureus COL
Staphylococcus aureus MRSA252
Staphylococcus aureus MSSA476
Staphylococcus aureus MW2
Staphylococcus aureus Mu50
Staphylococcus aureus N315
Staphylococcus epidermidis ATCC 12228
Streptococcus agalactiae III
Streptococcus agalactiae V
Streptococcus mutans
Streptococcus pneumoniae ATCC BAA-255
Streptococcus pneumoniae TIGR4
Streptococcus pyogenes MGAS10394
Streptococcus pyogenes MGAS315
Streptococcus pyogenes MGAS8232
Streptococcus pyogenes SF370
Streptococcus pyogenes SSI-1
Streptococcus thermophilus ATCC BAA-250
Streptococcus thermophilus CNRZ 1066
Streptomyces avermitilis
Streptomyces coelicolor
Symbiobacterium thermophilum
Synechococcus elongatus
Synechococcus sp. PCC 6301
Synechococcus sp. WH8102
Synechocystis sp.
Thermoanaerobacter tengcongensis
Thermotoga maritima
Thermus thermophilus HB27
Thermus thermophilus HB8
Treponema denticola
Treponema pallidum
Tropheryma whipplei TW08/27
Tropheryma whipplei Twist
Ureaplasma parvum
Vibrio cholerae
Vibrio parahaemolyticus
Vibrio vulnificus CMCP6
Vibrio vulnificus YJ016
Wigglesworthia glossinidia brevipalpis
Wolbachia pipientis wMel
Wolbachia sp.
Wolinella succinogenes
Xanthomonas axonopodis
Xanthomonas campestris campestris
Xanthomonas oryzae
Xylella fastidiosa 9a5c
Xylella fastidiosa Temecula1
Yersinia pestis 91001
Yersinia pestis CO-92
Yersinia pestis KIM5
Yersinia pseudotuberculosis
Zymomonas mobilis



Nearest 7 = Level 9 = All

Acinetobacter sp.
Aeropyrum pernix
Agrobacterium tumefaciens
Anabaena sp.
Anaplasma marginale
Aquifex aeolicus
Archaeoglobus fulgidus
Azoarcus sp.
Bacillus anthracis
Bacillus anthracis
Bacillus anthracis
Bacillus cereus
Bacillus cereus
Bacillus cereus
Bacillus clausii
Bacillus halodurans
Bacillus licheniformis
Bacillus subtilis
Bacillus thuringiensis
Bacteroides fragilis
Bacteroides thetaiotaomicron
Bartonella henselae
Bartonella quintana
Bdellovibrio bacteriovorus
Bifidobacterium longum
Bordetella bronchiseptica
Bordetella parapertussis
Bordetella pertussis
Borrelia burgdorferi
Borrelia garinii
Bradyrhizobium japonicum
Brucella melitensis
Brucella suis
Buchnera aphidicola
Buchnera aphidicola
Buchnera aphidicola
Burkholderia mallei
Burkholderia pseudomallei
Campylobacter jejuni
Campylobacter jejuni
Candidatus Blochmannia
Caulobacter crescentus
Chlamydia muridarum
Chlamydia pneumoniae
Chlamydia pneumoniae
Chlamydia pneumoniae
Chlamydia pneumoniae
Chlamydia trachomatis
Chlamydophila caviae
Chlorobium tepidum
Chromobacterium violaceum
Clostridium acetobutylicum
Clostridium perfringens
Clostridium tetani
Corynebacterium diphtheriae
Corynebacterium efficiens
Corynebacterium glutamicum
Coxiella burnetii
Deinococcus radiodurans
Desulfotalea psychrophila
Desulfovibrio vulgaris
Ehrlichia ruminantium
Ehrlichia ruminantium
Enterococcus faecalis
Erwinia carotovora
Escherichia coli
Escherichia coli
Escherichia coli
Escherichia coli
Francisella tularensis
Fusobacterium nucleatum
Geobacillus kaustophilus
Geobacter sulfurreducens
Gloeobacter violaceus
Gluconobacter oxydans
Haemophilus ducreyi
Haemophilus influenzae
Haloarcula marismortui
Halobacterium salinarium
Helicobacter hepaticus
Helicobacter pylori
Helicobacter pylori
Idiomarina loihiensis
Lactobacillus acidophilus
Lactobacillus johnsonii
Lactobacillus plantarum
Lactococcus lactis
Legionella pneumophila
Legionella pneumophila
Legionella pneumophila
Leifsonia xyli
Leptospira interrogans
Leptospira interrogans
Listeria innocua
Listeria monocytogenes
Listeria monocytogenes
Mannheimia succiniciproducens
Mesoplasma florum
Methanobacterium thermoautotrophicum
Methanococcus jannaschii
Methanococcus maripaludis
Methanopyrus kandleri
Methanosarcina acetivorans
Methanosarcina mazei
Methylococcus capsulatus
Mycobacterium bovis
Mycobacterium leprae
Mycobacterium paratuberculosis
Mycobacterium tuberculosis
Mycobacterium tuberculosis
Mycoplasma gallisepticum
Mycoplasma genitalium
Mycoplasma hyopneumoniae
Mycoplasma mobile
Mycoplasma mycoides
Mycoplasma penetrans
Mycoplasma pneumoniae
Mycoplasma pulmonis
Nanoarchaeum equitans
Neisseria meningitidis
Neisseria meningitidis
Nitrosomonas europaea
Nocardia farcinica
Oceanobacillus iheyensis
Onion yellows
Parachlamydia sp.
Pasteurella multocida
Photobacterium profundum
Photorhabdus luminescens
Picrophilus torridus
Porphyromonas gingivalis
Prochlorococcus marinus
Prochlorococcus marinus
Prochlorococcus marinus
Propionibacterium acnes
Pseudomonas aeruginosa
Pseudomonas putida
Pseudomonas syringae
Pyrobaculum aerophilum
Pyrococcus furiosus
Pyrococcus horikoshii
Ralstonia solanacearum
Rhizobium loti
Rhizobium meliloti
Rhodopirellula baltica
Rhodopseudomonas palustris
Rickettsia conorii
Rickettsia prowazekii
Rickettsia typhi
Salmonella paratyphi-a
Salmonella typhi
Salmonella typhi
Salmonella typhimurium
Shewanella oneidensis
Shigella flexneri
Shigella flexneri
Silicibacter pomeroyi
Staphylococcus aureus
Staphylococcus aureus
Staphylococcus aureus
Staphylococcus aureus
Staphylococcus aureus
Staphylococcus aureus
Staphylococcus epidermidis
Streptococcus agalactiae
Streptococcus agalactiae
Streptococcus mutans
Streptococcus pneumoniae
Streptococcus pneumoniae
Streptococcus pyogenes
Streptococcus pyogenes
Streptococcus pyogenes
Streptococcus pyogenes
Streptococcus pyogenes
Streptococcus thermophilus
Streptococcus thermophilus
Streptomyces avermitilis
Streptomyces coelicolor
Sulfolobus solfataricus
Sulfolobus tokodaii
Symbiobacterium thermophilum
Synechococcus elongatus
Synechococcus sp.
Synechococcus sp.
Synechocystis sp.
Thermoanaerobacter tengcongensis
Thermoplasma acidophilum
Thermoplasma volcanium
Thermotoga maritima
Thermus thermophilus
Thermus thermophilus
Treponema denticola
Treponema pallidum
Tropheryma whipplei
Tropheryma whipplei
Ureaplasma parvum
Vibrio cholerae
Vibrio parahaemolyticus
Vibrio vulnificus
Vibrio vulnificus
Wigglesworthia glossinidia
Wolbachia pipientis
Wolbachia sp.
Wolinella succinogenes
Xanthomonas axonopodis
Xanthomonas campestris
Xanthomonas oryzae
Xylella fastidiosa
Xylella fastidiosa
Yersinia pestis
Yersinia pestis
Yersinia pestis
Yersinia pseudotuberculosis
Zymomonas mobilis

Nearest 7 = Level 9 = All

Supplementary figures

The next pages contain a list of supplementary figures providing additional insight on the approaches used to predict protein interactions using coevolution. All these figures are referenced in the main manuscript but relegated to the backend to improve the readability.

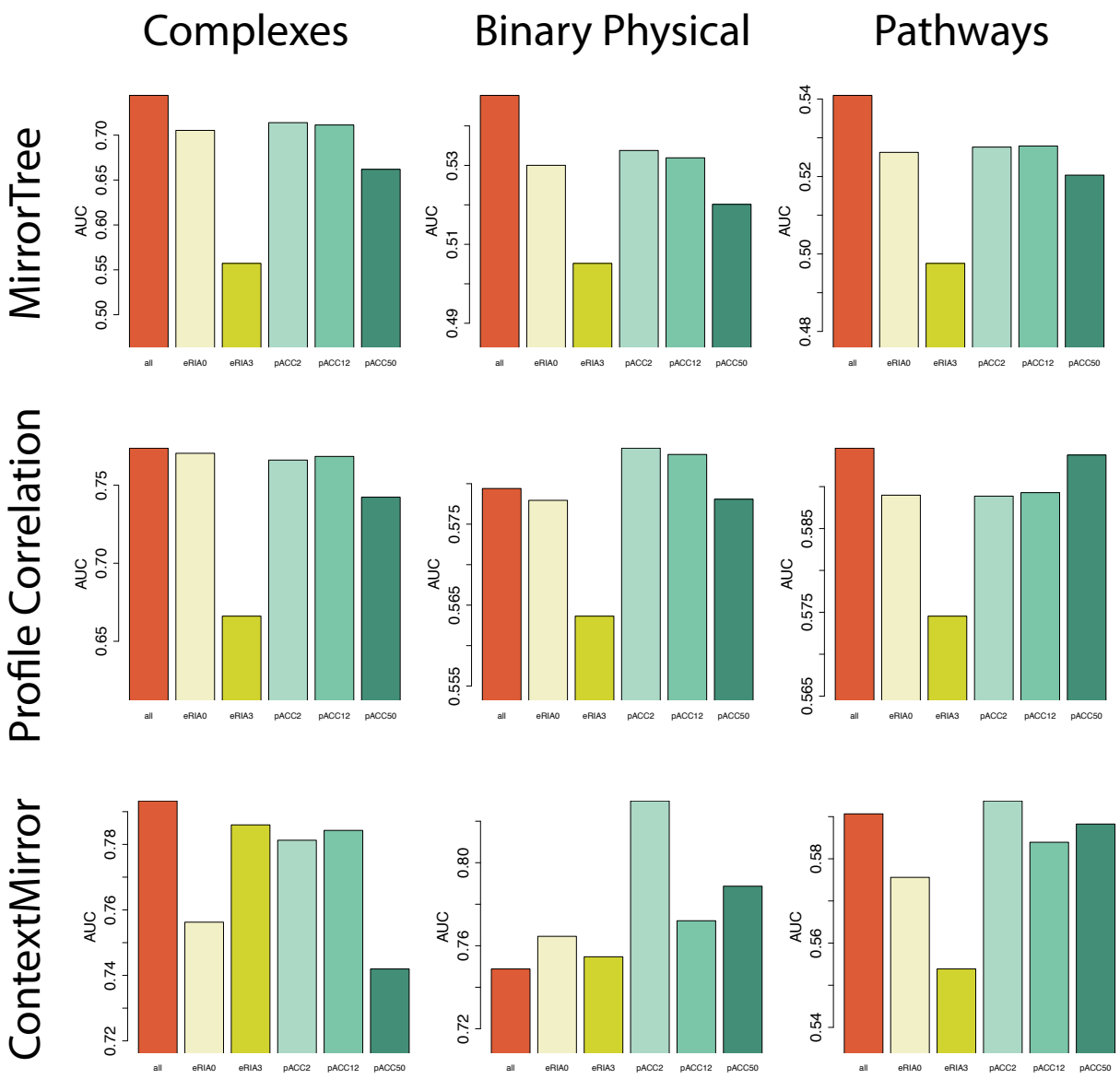


Figure S1: **Performance obtained using different combinations of: phylogenetic tree comparative methods, interaction evidence and predicted accessibility filter**. Different performances are calculated using the Area Under the ROC Curve (AUC). In order to highlight the differences between the different methods and interaction datasets, the scales were adjusted independently.

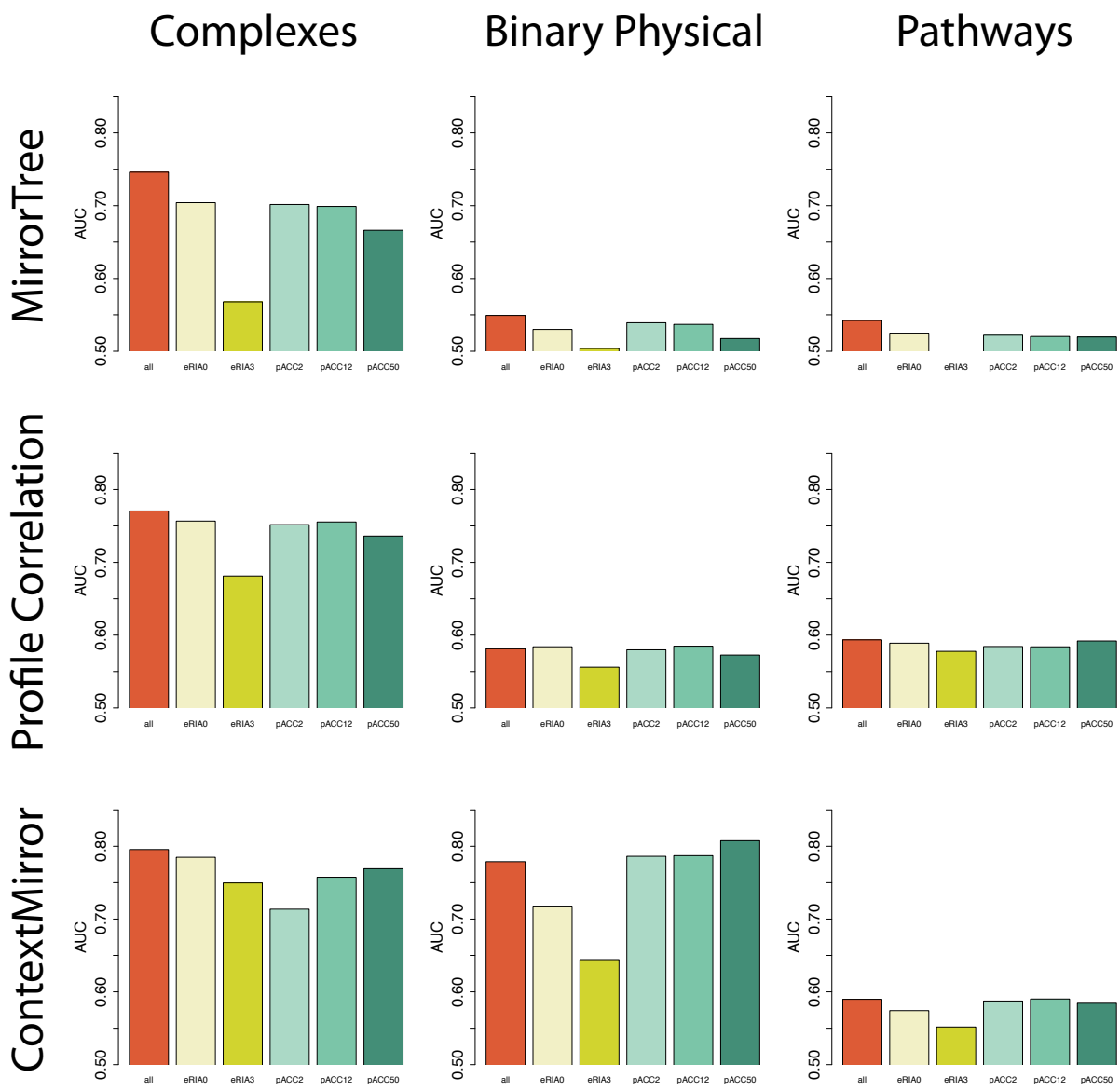


Figure S2: **Performances of the different methods predicting different types of interactions using trees derived from positions with different predicted accessibility features**. The performance is evaluated as the Area Under the ROC Curve (AUC) using predicted accessibility derived from MSAs of orthologs.

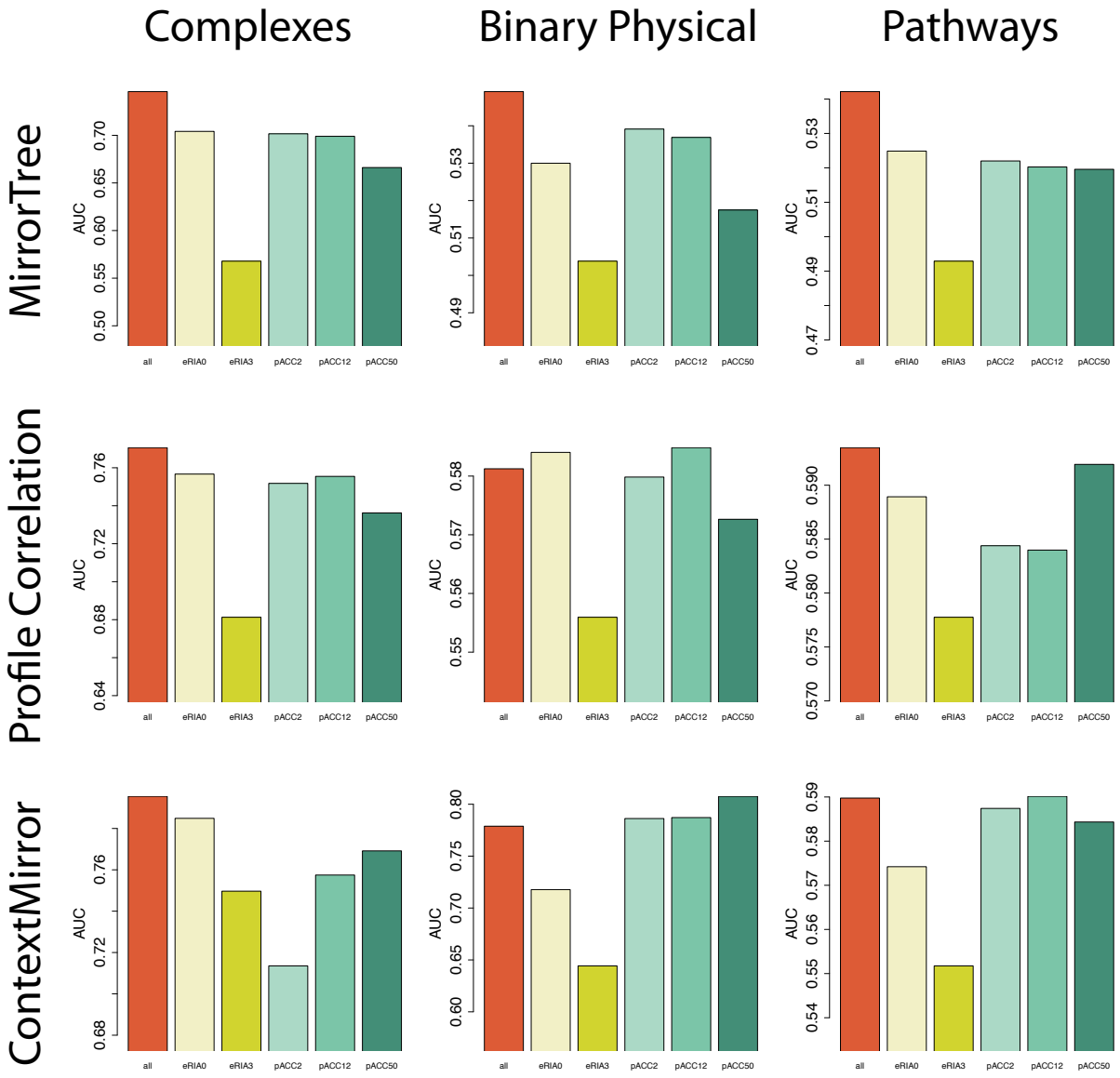


Figure S3: **Performances of the different methods predicting different types of interactions using trees derived from positions with different predicted accessibility features.** The performance is evaluated as the Area Under the ROC Curve (AUC) using predicted accessibility derived from MSAs of orthologs. In order to highlight the differences, the axis scales were adjusted independently for each case.

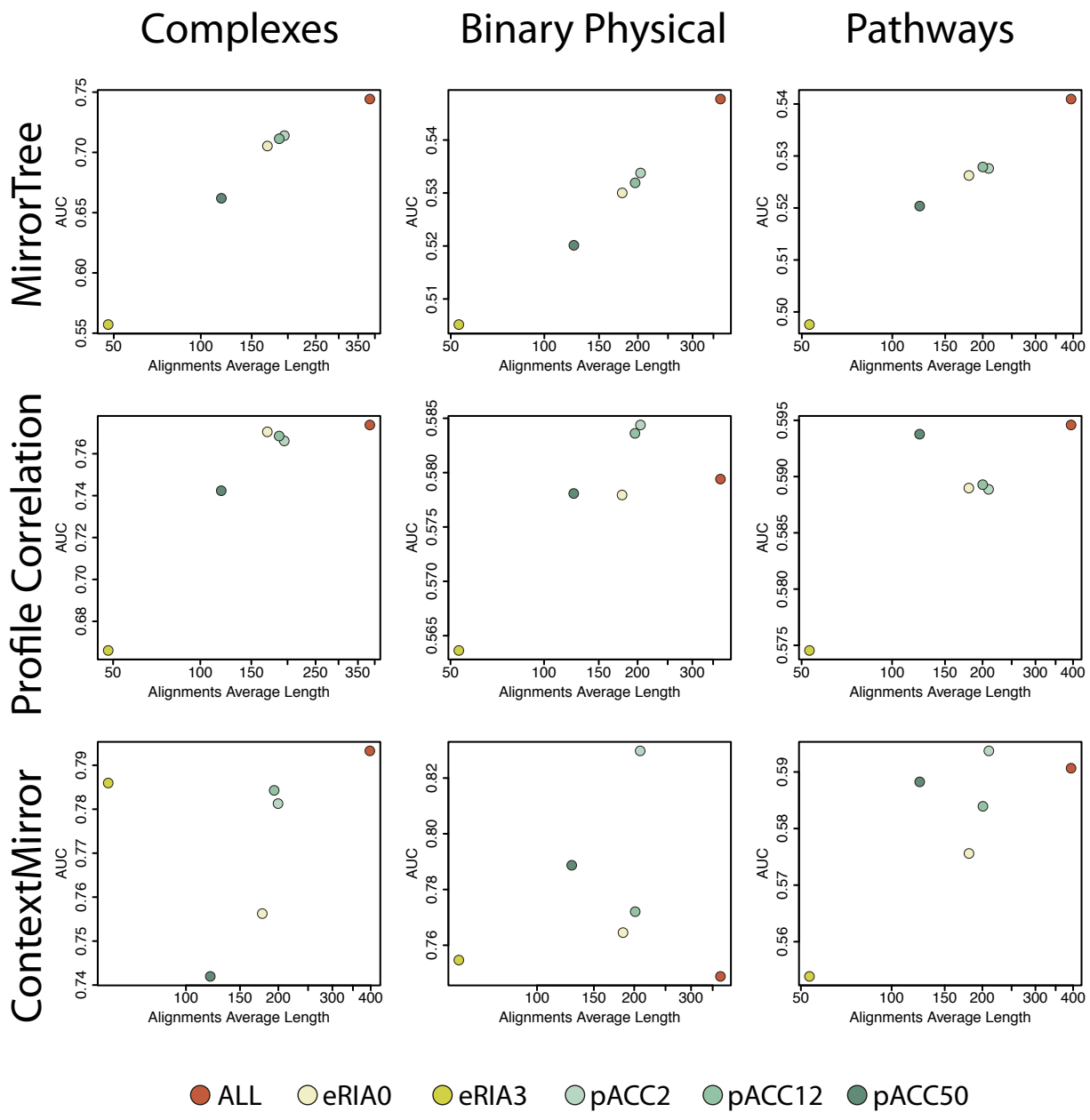


Figure S4: **Relationship between the performances of the different methods and the lengths of the virtual alignments for the different datasets**. The length of the virtual alignment is the number of positions (fulfilling a given predicted accessibility criteria –colors-) used to construct the trees.

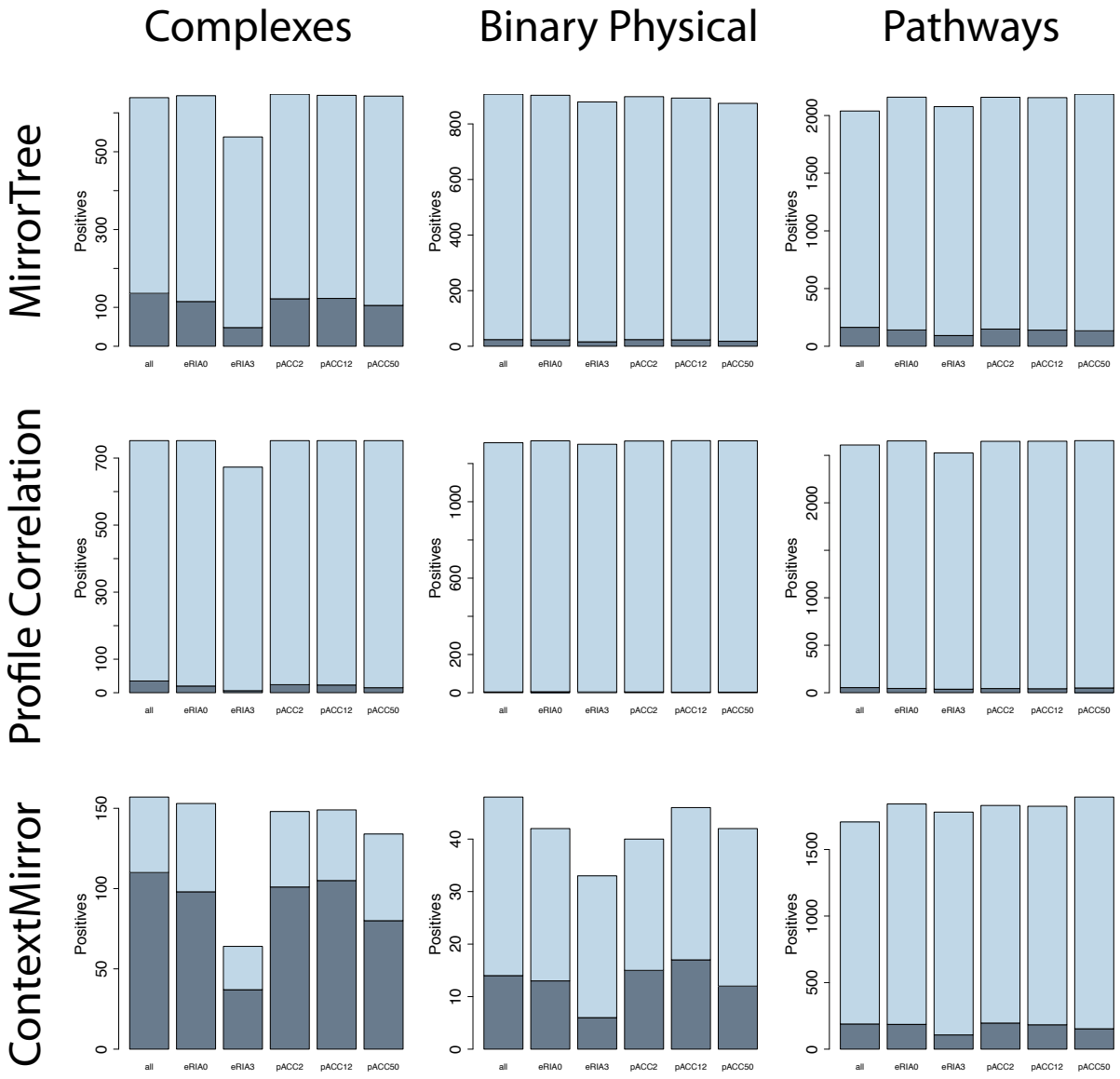


Figure S5: **Positive predictions obtained using different combinations of: phylogenetic tree comparative methods, interaction evidence and predicted accessibility filter**. The bars represent the total number of positives (nP) for which the calculation could be done (fulfilling organisms in common and *P*-value criteria) for a given prediction. The dark-blue bars, represent the subset of true positives among the first nP protein pairs.

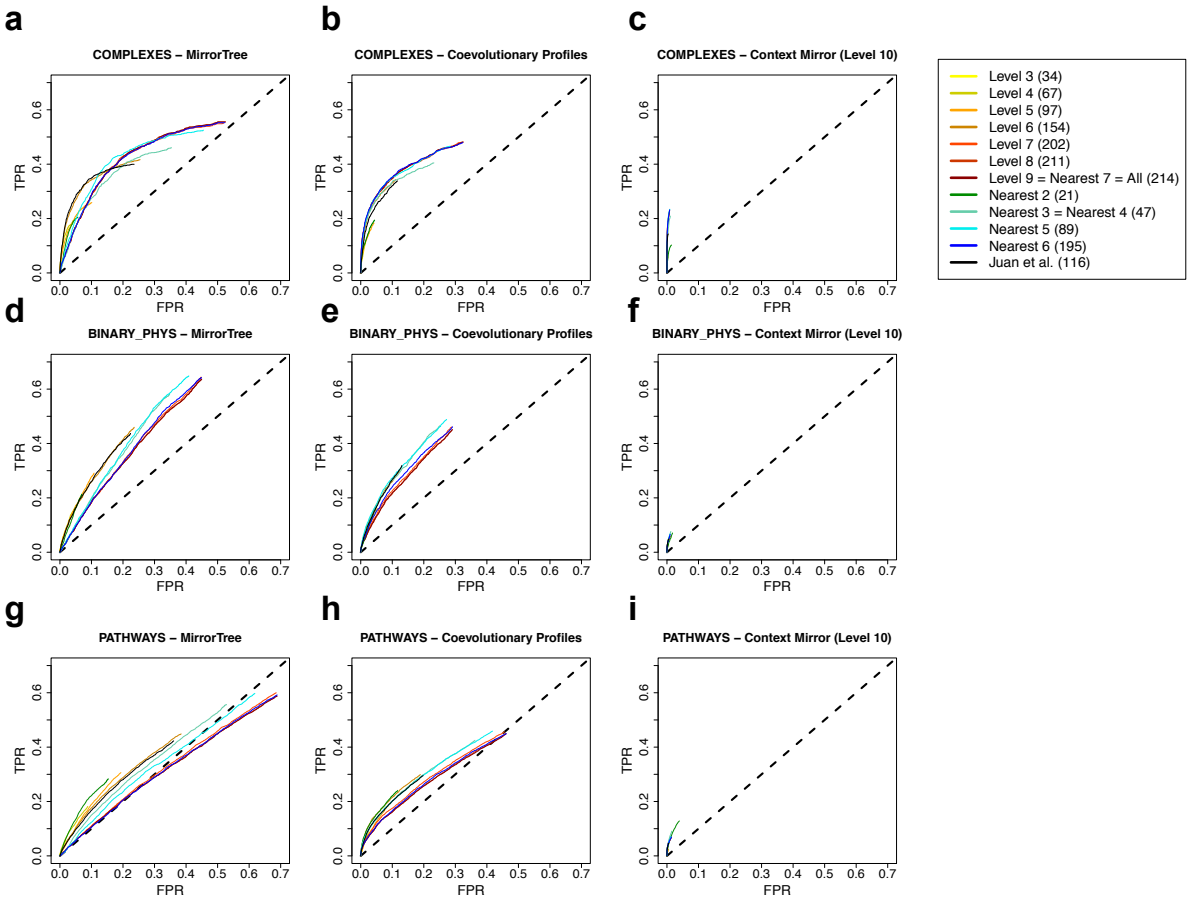


Figure S6: **Matrix of partial ROC curves**. The partial ROC curves evaluate the performance of a given list of predictions obtained by the combination of a methodology (columns), a dataset of interactions (rows) and a set of organisms (colors according to the legend). In the legend, the number of organisms present in the dataset is included within brackets. The dashed line represents the performance of a random classifier. The different plots are in the same scale in order to compare their global performances.

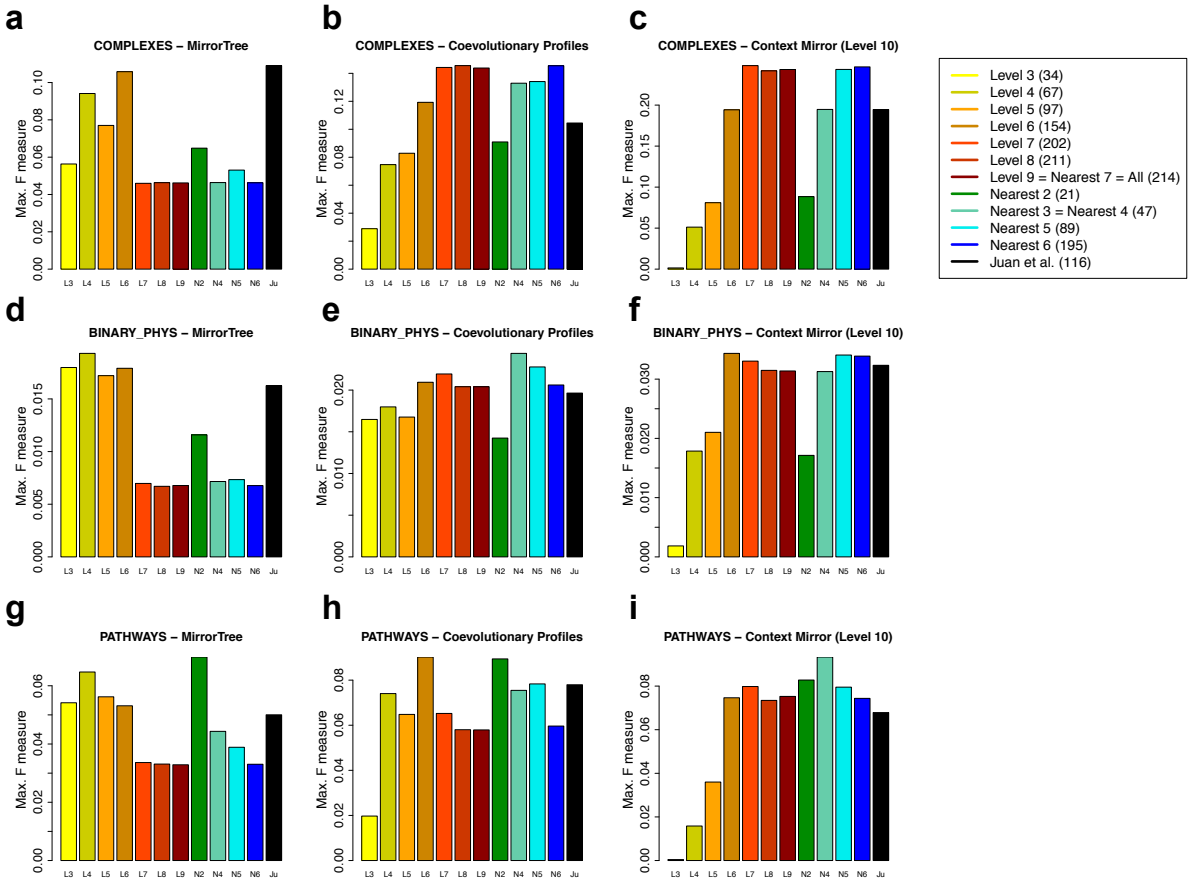


Figure S7: **Maximum F-measure for each method/organism set**. The optimal F-measure along the range of possible cutoffs is showed in these bar plots. The rows represent the interaction dataset and the columns the methods. For a given combination of method and interaction dataset the colored bars represent different sets of organisms used to reconstruct the phylogenetic trees. Extended labels, as well as the number of organisms are shown in the legend.

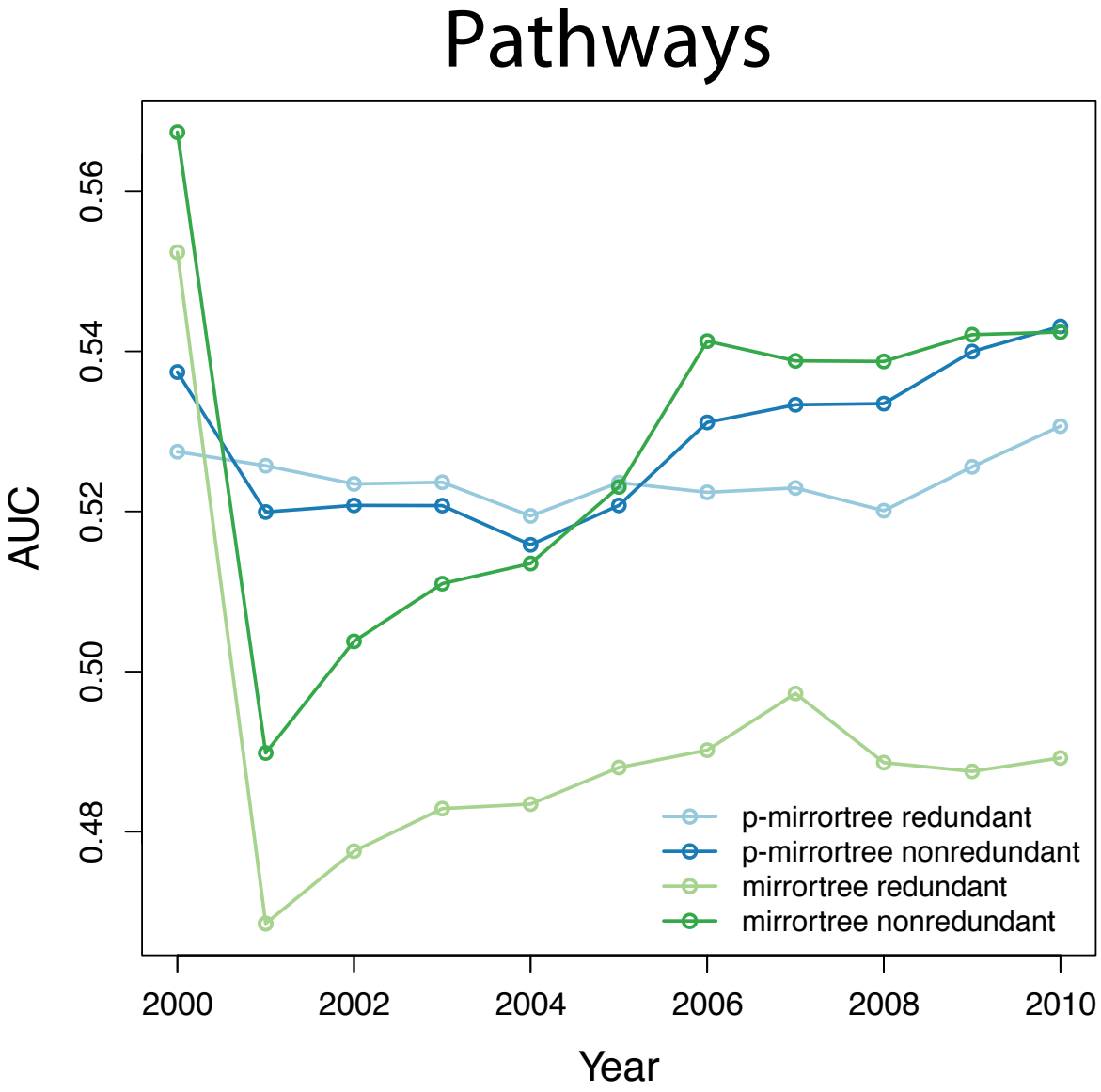


Figure S8: **Performance of the *mirrortree* and *p-mirrortree* methods when predicting interactions using different sets of organisms based on the fully-sequenced genomes available in the period 2000-2010 and different taxonomical redundancies.** The performances were evaluated in terms of AUC using a gold standard dataset of protein interactions defined as co-membership in the same KEGG pathway.

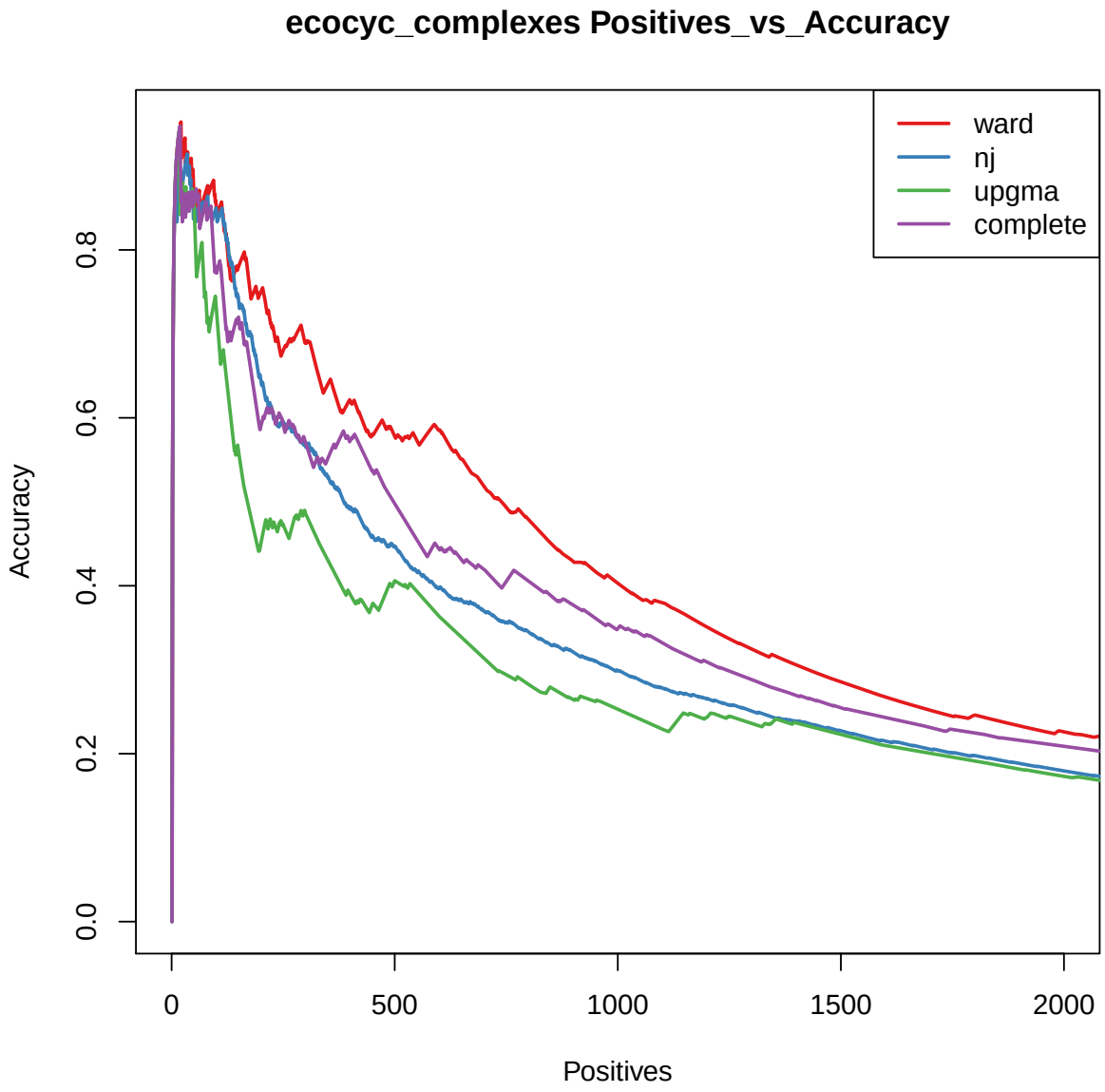


Figure S9: **Accuracy *vs.* number of positives in 4 different hierarchichal clustering algorithms predicting protein interactions**. Cophenetic distances from the resulting clustering of coevolutionary profiles were used to score the predictions. The clusterings algorithms were based on Ward's minimum variance ("ward"), neighbor-joining ("nj"), UPGMA ("upgma") and complete linkage ("complete"). Protein interactions were evaluated using the "Complexes" gold standard dataset.

Published works

The next pages contain three co-authored studies which partly overlap with the different topics of this thesis.

BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 10 2010, pages 1370–1371
doi:10.1093/bioinformatics/btq137

Sequence analysis Advance Access publication March 30, 2010

Studying the co-evolution of protein families with the Mirrortree web server

David Ochoa and Florencio Pazos*

National Centre for Biotechnology, Computational Systems Biology Group (CNB-CSIC), c/ Darwin, 3. Cantoblanco, 28049 Madrid, Spain

Associate Editor: Burkhard Rost

ABSTRACT

Summary: The Mirrortree server allows to graphically and interactively study the co-evolution of two protein families, and investigate their possible interactions and functional relationships in a taxonomic context. The server includes the possibility of starting from single sequences and hence it can be used by non-expert users.

Availability and Implementation: The web server is freely available at http://csbg.cnb.csic.es/mtserver. It was tested in the main web browsers. Adobe Flash Player is required at the client side to perform the interactive assessment of co-evolution.

Contact: pazos@cnb.csic.es

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on January 21, 2010; revised on March 23, 2010; accepted on March 26, 2010

1 INTRODUCTION

A lot of biological knowledge is hidden in the complex networks of relationships of different nature between molecular entities. In the case of proteins, their biological roles can only be fully understood in the context of their interaction with others. This importance in deciphering as much as possible of the complex network of interactions and functional relationships between proteins has led to the development of specific experimental (Shoemaker and Panchenko, 2007a) and computational (Shoemaker and Panchenko, 2007b) techniques for this task. One family of these computational techniques is based on the observed relationship between protein interactions and co-evolution [similarity of evolutionary histories as represented by phylogenetic trees; see Pazos and Valencia (2008) and references herein]. This approach, termed *mirrortree*, has been applied not only to look for interaction partners in large datasets of proteins (e.g. Juan *et al*., 2008), but also to study in depth the co-evolution and interactions in particular pairs of protein families (e.g. Dou *et al*., 2006; Labedan *et al*., 2004; McPartland *et al*., 2007). Many authors developed variations and different implementations of this approach [e.g. see references in Pazos and Valencia (2008)], but none of them are intended to be operated by non-experts users. They are either very specific for certain needs or are distributed as non-interactive command-line programs or require a complex preparation of the input data (e.g. generation of the multiple sequence alignments (MSAs) and/or phylogenetic trees). This precludes these techniques from being used by most molecular biologists.

In this work, we present the Mirrortree server, an automatic system for the interactive assessment of co-evolutionary features between two protein families. The system only requires as input the sequence of a single representative of each family to start, which allows it to be used by non-bioinformaticians. All the subsequent steps (search for homologues, localization of orthologues, generation and filtering of MSAs and trees, and tree comparison) are fully automatic. Nevertheless, expert users have the possibility of providing their (manually curated) MSAs or trees. Moreover, the tree comparison is done in an interactive interface that allows users to study in depth the co-evolution of their families and investigate their interactions in a taxonomic context.

2 WORKFLOW

Supplementary Material 1 contains an exhaustive description of the server workflow. What follows is a short description. Each one of the two input sequences is BLASTed (Altschul *et al*., 1997) against the Integr8 database of fully sequenced genomes (Kersey *et al*., 2005). The list of putative homologues is filtered to discard fragments, divergent sequences, etc. The remaining sequences are aligned with Muscle (Edgar, 2004). The resulting MSA is filtered again (see Supplementary Material 1 for details) and only one homologue per species is retained as the putative orthologue (the one with highest similarity to the master). The final MSA of putative orthologues is used to construct a phylogenetic tree with the 'neighbour-joining' (NJ) algorithm implemented in ClustalW (Chenna *et al*., 2003). Expert users can bypass these steps by providing their own MSAs or phylogenetic trees (i.e. generated with more sophisticated techniques than NJ). The computationally expensive steps are delegated to a computer cluster. As an example, running the whole process for two families of around 800 residues long with 120 species in common takes 10 min.

3 INTERFACE

When the process is completed, the user receives an e-mail containing a link to the interactive Flash-based visualization of the trees of the two families (Fig. 1), as well as files with useful intermediate results (MSAs and trees for the two families, static graphical representations of the mirroring trees, etc). Organisms present in both families are connected by lines in this representation. Tree branches can be swapped in order to confront matching clades between the two trees and obtain a better representation. The tree

*To whom correspondence should be addressed.

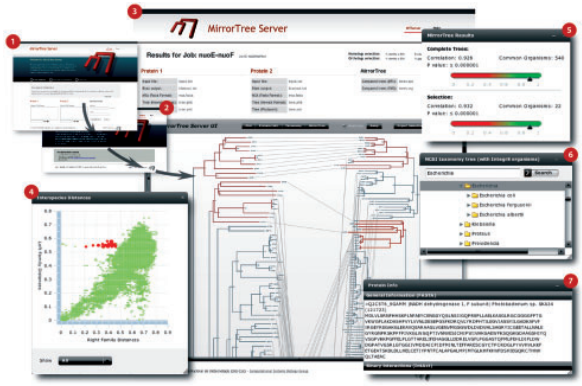


Fig. 1. Interface of the Mirrortree server. (**1**) Job submission page. (**2**) Job ID and status. (**3**) Main interface for viewing and manipulating the trees. The different panels can be shown/hidden and freely moved/resized in a windows-like manner. (**4**) Panel with the distance correlation plot. (**5**) Tree and sub-tree similarity scales and associated *P*-values. (**6**) Taxonomy browser. (**7**) Uniprot information for individual proteins.

representation can be zoomed and the user can select different proteins (leaves) or whole clades (internal nodes) in both trees in order to restrict the calculation of tree similarity to certain groups of organisms. Panels with additional tools and information are arranged on the top of this representation and can be shown/hidden and freely moved/resized in a windows-based interface (Fig. 1). One of these panels shows the similarity of the trees as calculated by *mirrortree* in a colour scale. The tree similarity for the current selection is also shown in this panel. Another panel shows information available for the selected proteins (leaves) in the Uniprot resource (Uniprot Consortium, 2009), such as protein name, sequence, organism and reported interactions. Organism selection can also be done by taxonomic criteria using the included taxonomy browser (Fig. 1), i.e. to evaluate the co-evolution in a certain kingdom or family. Selections in the tree are also shown in the taxonomy browser. The sub-alignment for the sequences in the current selection can be exported for further analysis. Finally a plot with a simplified representation of the correlation between the inter-protein distances in both families is also shown. This plot can show all the distances or only the ones involving the selected organisms. This plot is very useful to detect outliers: clouds of points far from the diagonal representing non-correlated distances that decrease the overall similarity of the trees. In many cases, these are related to non-standard evolutionary events such as horizontal gene transfer (Pazos *et al*., 2005). Selections of points in this plot cause the corresponding organisms/clades in the trees to be selected. The server has many other features extensively explained in a help file. There is also a guided tutorial for illustrating the kind of studies that can be performed with the server.

4 CONCLUSION

The Mirrortree server is the first system for interactively assessing the co-evolution between two protein families in order to evaluate their possible interactions in a taxonomic framework. There are related systems such as TSEMA (Izarzugaza *et al*., 2008) which, based on the same relationship between protein interactions and tree similarity, are nevertheless intended for predicting the mapping (connections between the leaves) between two families already known to interact. Moreover, that server does not include the possibility of automatically generating MSAs and hence it is more difficult to be used by non-experts.

An important requirement for a computational tool to be used by biologists is simplicity. That left most existing tools for studying co-evolution and predicting protein interactions out of their standard toolkit. The Mirrortree server was developed with the goal of being amenable to be used by non-experts, in such a way that any user can interactively study the co-evolution between his/her families of interest in a taxonomic context starting with single sequences.

ACKNOWLEDGEMENTS

We want to thank Octavio Diaz-Pines and the members of the CTI-CSIC for computer support. We also want to acknowledge Daniel Lopez (CNB-CSIC), David Juan and Alfonso Valencia (CNIO) for comments and suggestions.

Funding: BIO2006-15318 project of the Spanish Ministry for Science and Innovation (in part); PhD fellowship of the Basque Country Government (to D.O.).

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Chenna,R. *et al*. (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.*, **31**, 3497–3500.

Dou,T. *et al*. (2006) Co-evolutionary analysis of insulin/insulin like growth factor 1 signal pathway in vertebrate species. *Front. Biosci.*, **11**, 380–388.

Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.

Izarzugaza,J.M. *et al*. (2008) Enhancing the prediction of protein pairings between interacting families using orthology information. *BMC Bioinformatics*, **9**, 35.

Juan,D. *et al*. (2008) High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. *Proc. Natl Acad. Sci. USA*, **105**, 934–939.

Kersey,P. *et al*. (2005) Integr8 and Genome Reviews: integrated views of complete genomes and proteomes. *Nucleic Acids Res.*, **33**, D297–D302.

Labedan,B. *et al*. (2004) Using quaternary structures to assess the evolutionary history of proteins: the case of the aspartate carbamoyltransferase. *Mol. Biol. Evol.*, **21**, 364–373.

McPartland,J.M. *et al*. (2007) Coevolution between cannabinoid receptors and endocannabinoid ligands. *Gene*, **397**, 126–135.

Pazos,F. *et al*. (2005) Assessing Protein co-evolution in the context of the tree of life assists in the prediction of the interactome. *J. Mol. Biol.*, **352**, 1002–1015.

Pazos,F. and Valencia,A. (2008) Protein co-evolution, co-adaptation and interactions. *EMBO J.*, **27**, 2648–2655.

Shoemaker,B.A. and Panchenko,A.R. (2007a) Deciphering protein-protein interactions. Part I. Experimental techniques and databases. *PLoS Comput. Biol.*, **3**, e42.

Shoemaker,B.A. and Panchenko,A.R. (2007b) Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners. *PLoS Comput. Biol.*, **3**, e43.

Uniprot Consortium. (2009) The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res.*, **37**, D169–D174.

Herman *et al. BMC Bioinformatics* 2011, **12**:363
http://www.biomedcentral.com/1471-2105/12/363

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Selection of organisms for the co-evolution-based study of protein interactions

Dorota Herman[1,3], David Ochoa[1], David Juan[2], Daniel Lopez[1], Alfonso Valencia[2] and Florencio Pazos[1*]

Abstract

Background: The prediction and study of protein interactions and functional relationships based on similarity of phylogenetic trees, exemplified by the *mirrortree* and related methodologies, is being widely used. Although dependence between the performance of these methods and the set of organisms used to build the trees was suspected, so far nobody assessed it in an exhaustive way, and, in general, previous works used as many organisms as possible. In this work we asses the effect of using different sets of organism (chosen according with various phylogenetic criteria) on the performance of this methodology in detecting protein interactions of different nature.

Results: We show that the performance of three *mirrortree*-related methodologies depends on the set of organisms used for building the trees, and it is not always directly related to the number of organisms in a simple way. Certain subsets of organisms seem to be more suitable for the predictions of certain types of interactions. This relationship between type of interaction and optimal set of organism for detecting them makes sense in the light of the phylogenetic distribution of the organisms and the nature of the interactions.

Conclusions: In order to obtain an optimal performance when predicting protein interactions, it is recommended to use different sets of organisms depending on the available computational resources and data, as well as the type of interactions of interest.

Background

There are many computational methods for predicting protein interactions and functional relationships (see [1-3] for recent reviews). Among them, two types of techniques, "phylogenetic profiling" and "similarity of phylogenetic trees", are based on the fact that interacting or functionally related proteins are co-evolving at different levels, defining co-evolution as interdependence between evolutionary histories [4,5].

Phylogenetic profiling [6] is based on the intuitive idea that the genes of two functionally related protein families, which need each other to work, will tend to be both present in the same set of organisms, and probably absent together in the complementary set. A "phylogenetic profile" is a vector representing the pattern of presence/absence of a given gene in a set of organisms, eventually with quantitative information on the sequence similarity of the genes respect to that in a reference organism [7]. Similarity between two of these vectors has been shown to be a good indicator of functional relationship between the families they represent. The similarity of presence/absence patterns between interacting proteins can be seen as a reflection of an extreme case of evolutionary dependence (co-evolution) since the "existence" of the proteins themselves depends on each other.

Co-evolution between interacting or functionally related protein families is also reflected in their phylogenetic trees, being these more similar than expected. Such similarity was first qualitatively evaluated and latter quantified for large collections of interacting and non-interacting protein pairs in order to statistically assess its relationship with interaction [8]. Since then, this idea was applied to study many interacting families, and many groups developed different implementations and variations of the methodology (see [2,4,5] for recent reviews). The basic *mirrortree* methodology for predicting whether two proteins of a given organism interact or not starts by looking for orthologs of these two sequences in a set of genomes. Multiple sequence

* Correspondence: pazos@cnb.csic.es
[1]Computational Systems Biology Group, National Centre for Biotechnology (CNB-CSIC), C/Darwin, 3, Cantoblanco, 28049 Madrid, Spain
Full list of author information is available at the end of the article

Herman et al. BMC Bioinformatics 2011, 12:363
http://www.biomedcentral.com/1471-2105/12/363

alignments are then generated for these two sets of orthologs and phylogenetic trees are obtained from them. Pairwise distances are then calculated for all possible pairs of sequences in both sets. Finally, the similarity between these two sets of distances is evaluated with a linear correlation coefficient, using only the distances involving organisms present in both sets. A high correlation coefficient is indicative of similar trees and hence of possible co-evolution. This co-evolutive trend points to a possible interaction or functional relationship between the proteins. This methodology has recently been fully automated and implemented in a web server which allows non-expert users to apply it starting with single sequences [9]. Moreover, this basic methodology has been improved in many ways by different authors (see [4,5] for recent reviews). For example, the background similarity expected between any pair of trees due to the underlying speciation process has been subtracted in different ways in order to improve the predictions [10-12]. More recently, networks representing the pair-wise tree similarities for all proteins in a given genome have been used to improve the prediction of interaction partners and to get insight into the substructure and functioning of macromolecular complexes [13]. Part of this last methodology consist on representing the co-evolutionary context of a given protein by a vector containing its tree similarities (correlation values) with the rest of the proteins, and then re-evaluating the eventual co-evolution between two proteins as the correlation between their corresponding vectors (co-evolutionary profiles). In the same framework of genome-wide co-evolutions, a partial correlation study allows to separate specific from non-specific co-evolutions [13]. It has been shown that these two variants are better predictors of interaction than the original tree correlations.

Both *mirrortree* and "phylogenetic profiling" use a reference set of organisms for looking for orthologs and building the phylogenetic trees or presence/absence profiles respectively. The characteristics of this set (number of organisms, phylogenetic distribution, etc.) are expected to influence the performance of these methodologies. For phylogenetic profiling, some pioneering studies addressed this problem by evaluating the effect of this reference set of genomes on the performance and range of applicability of the methodology [14,15]. Nevertheless, no equivalent study has been done for *mirrortree* and related methodologies. In most studies, the authors use all genomes available in a given resource/database (see references in [4,5]) or, in some cases, they remove redundancy at the strain level [13]. It is worth studying the effect of the organism set in the performance of the *mirrortree*-related methodologies for three main reasons: i) There could be a subset of organisms yielding better results than the whole set of available genomes; ii) different types of interactions (physical, functional, ...) could be better detected using different subsets of organisms; and iii) with the growing number of completely-sequenced genomes, there will be a point in the future were it would not be possible to use all. In such case, it would be valuable to have "recipes" on which subset(s) to use, phrased in terms of number of organisms, phylogenetic distribution, etc.

In this work we explore the effect of using different reference sets of organisms in the performance of the original *mirrortree* algorithm [8] and two of its more recent variants: *profile-correlation* and *context-mirror* [13]. Starting with the set of 214 genomes used by Juan et al. [13], we took different subsets sampled according with different taxonomic criteria, and evaluate the performance of these methodologies using as gold standards sets of interactions of different nature (physical, functional, ...). Our goal is to get insight on the influence of these factors on the co-evolutionary analyses. The results obtained allowed us to propose a number of pragmatic recipes for the use of these methodologies in terms of which subset is better for detecting each particular type of interactions, and which subset to use when the number of available sequenced genomes makes it impossible to use all. Apart from the results obtained from a large scale evaluation, we also show particular examples to illustrate how using different sets of organisms can drastically affect the observed co-evolution between proteins.

Methods

For comparative purposes, we used as initial set of organisms all the Eubacteria and Archaea that were fully sequenced and available in the integr8 database [16] at the time when Juan *et al.* work was performed: 214 genomes. (In that work, redundancy was removed in order not to include very similar organisms, ending up in a final set of 116 organisms.) We then sampled this initial set according with different taxonomic criteria using *E coli K12* as reference organism, and evaluated the performance of three *mirrortree*-related methodologies in a number of sets representing different types of interactions and using these sampled subsets of organisms as reference sets. Figure 1 illustrates the process.

Selection of different subsets of organisms

We used the NCBI taxonomic tree [17] as framework for the taxonomy-based selection of organisms. This tree classifies organisms according with a pre-defined hierarchy in which the root represents "cellular organisms", the first level represents the "superkingdoms" (Archaea and Eubacteria in our dataset, which does not include eukarya), the next one the "phylums", and so on. This tree does not contain quantitative information

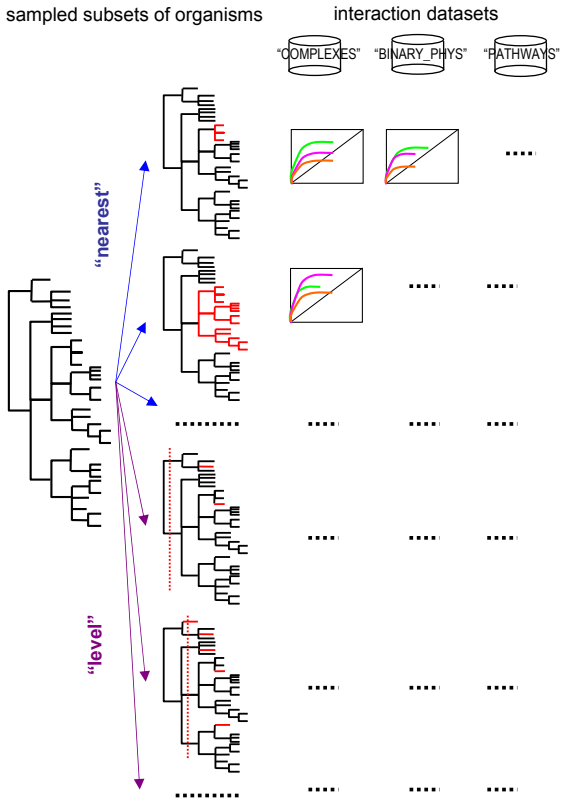


Figure 1 Schema of the methodology.From an initial set of organisms with completely sequenced genomes (left), a number of subsets (red) are constructed according with two taxonomic criteria: "nearest" (blue) - following the taxonomy of the reference organism (*E coli K12*) back to the root of the taxonomic tree, all the genomes belonging to each node visited (*E coli* species, Enterobacteriaceae family, etc.) are taken; "level" (purple) - the tree is successively cut at each taxonomic level (superkingdom, phylum, ...) and one organism is taken from each one of the resulting groups (the one with the largest proteome). On the other hand, a number of "gold standard" interaction datasets representing physical and functional interactions of different nature are used (top). For each combination interaction dataset/organism subset, the performance of the three mirrortree-based methodologies is assessed with a partial-ROC analysis (colored curves).

on phylogenetic distances between organisms. Two criteria were used for performing the selections:

- "Nearest". Starting from our reference organism (*E coli K12*) we follow its taxonomy back to the root of the tree and successively take all the organisms belonging to each node. So "nearest_1" represents the *E coli* species (4 organisms -strains-), "nearest_2" contains the Enterobacteriaceae family (21 organisms), and so on up to "nearest_6" which represents the Bacteria superkingdom (195 organisms) and "nearest_7" (whole dataset, bacteria+archaea, 214 organisms). Four organisms are represented in the trees but not used due to the lack of information on their proteomes in the NCBI data. This sampling is designed to evaluate the effect of including close vs. distant organisms in the performance, as well as the effect of the redundancy to some extent (Figure 1).
- "Level". The taxonomic tree is successively cut at each level of the hierarchical classification starting from the root (superkingdom, phylum, ...) and one organism is taken from each resulting group. The criterion for selecting an organism within a group is simply to use that with the highest number of proteins in its genome. The rationale for doing this is to maximize the chances of finding orthologs in that genome in subsequent steps of the process. So, "level_1" would contain 2 organisms (one eubacteria and one archaea), "level_2" contains 16 organisms, one for each phylum. And so on up to "level_9" which represents the whole dataset (214 genomes). This experiment is designed to quantify the effect of sampling homogeneously the taxonomy at different levels of granularity (Figure 1).
- For comparative purposes we also included the set of genomes used in Juan *et al* [13] (116 genomes). This set is very similar to our "level_5" (97 organisms).

Due to the requirement of 15 or more organisms in common between the trees of two protein families (see next point), some of these subsets are never used in practice. The lists of organisms in the final 12 subsets used, as well as representations of their taxonomic distributions, are available in the "Additional file s1".

Datasets of protein interactions and functional relationships

We used as gold standards to asses the methods' performance three datasets representing *E. coli* protein interactions of different nature and with different characteristics and peculiarities. "PATHWAYS": Functional interactions inferred as co-presence in metabolic pathways taken from the EcoCyc resource [18]. This dataset comprises 4,491 pairs between 719 proteins. "COMPLEXES": Physical interactions (not necessarily direct) inferred by co-presence in macromolecular complexes experimentally determined and taken also from EcoCyc (1,354 pairs between 591 proteins). "BINARY_PHYS": Physical direct binary interactions obtained from the MPIDB database [19]. These have been manually curated from the literature or imported from other databases, providing a high-confidence gold standard to evaluate putative physical direct interactions. The version we used of this database contains 2,103 binary interactions between 1,538 different *E. coli* proteins. The

Herman et al. BMC Bioinformatics 2011, 12:363
http://www.biomedcentral.com/1471-2105/12/363

first two datasets were previously used by Juan and co-workers [13], while the last is used here for the first time.

For each dataset, a set of negative examples (proteins assumed not to interact physically or functionally) is constructed by generating all possible pairs between the proteins involved in the positive (interacting) pairs.

Co-evolution-based prediction of protein interactions

We applied three methods used in Juan *et al.* [13] to predict interacting pairs of proteins using the different sets of reference genomes discussed above for constructing the trees.

The starting point for all methodologies is the generation of phylogenetic trees of orthologs for all *E coli* proteins using the reference sets of organisms sampled as described above. For detecting the ortholog of a given *E coli* protein in each genome we used the "BLAST best bi-directional hit" criterion, with an E-value cutoff of 10E-5, and requiring an alignment coverage of 70%. The orthologs found are aligned with Muscle [20] using the default parameters of this program. Then, a phylogenetic tree is generated from this alignment using the neighbor-joining algorithm implemented in ClustalW [21], excluding the gaps for the distance calculation. A matrix containing the pair-wise distances between all orthologs is generated from this tree by summing the lengths of the branches separating the corresponding leaves.

The *mirrortree* method (MT) evaluates the co-evolution between two proteins by calculating the linear correlation coefficient between their corresponding distance matrices. A minimum of 15 species in common between their trees is required for evaluating a given pair. Only correlation values supported by a (tabulated) P-value of 10E-5 or lower are considered.

A matrix containing the significant pair-wise tree correlations (P-value ≤ 10E-5) for all pairs of proteins within the genome of *E coli* is used as input for the *profile-correlation* method (PC). A row (or column) in this matrix (co-evolutionary profile) contains the correlations between a given protein and all the others in the genome, and can be considered as a representation of the co-evolutionary context for that protein. The *profile-correlation* method re-assesses the co-evolution between two proteins by calculating the linear correlation between their respective co-evolutionary profiles. Finally, the *context-mirror* method (CM) assesses the influence of third proteins in a given co-evolutionary signal observed for two proteins using a partial correlation criterion. This allows separating specific co-evolution (particular to a given pair of proteins) from general co-evolutionary trends involving many proteins. So this method produces results at different "levels" of specificity. See [13] for a more detailed description of these methodologies.

Evaluation

For each pair of proteins in the *E coli* genome fulfilling the requirements mentioned above, we have the scores of the three methods (*mirrortree*, *profile correlation* and *context-mirror*) based on a given sampled subset of organisms. As commented above, for *context-mirro*r the results are split in different levels of co-evolutionary specificity. In addition, we know whether that pair represents a true interaction or functional relationship according with the datasets described earlier. So, for each combination method/dataset/subset of organisms we have a large list of protein pairs sorted by the score of the method, being each pair labeled as "positive" (the two proteins interact according with the dataset) or "negative" (the two proteins are assumed not to interact).

We apply "receiver operating characteristic" analysis (ROC) [22] to these lists to assess the capacity of the method to separate the positives from the negatives. For each of these lists, the ROC analysis generates a plot of "true positives rate" (TPR) against "false positives rate" (FPR) when varying the classification threshold (score of the method). Curves above the diagonal in this plot represent methods with some discriminative power, being this discriminative capacity better as the curve gets closer to the top-left corner of the plot. Due to the requirement of 15 or more organisms in common in order to evaluate a given pair, the same method applied to the same interaction dataset can produce lists with very different number of pairs (both negatives and positives) when based on different subsets of organisms (trees with different number of leaves). In order to compare the ROC curves in these cases, FPR's and TPR's are calculated respect to the total number of pairs (positives and negatives) in the original dataset, and not respect to the number of pairs rendered by a given subset of organisms. Moreover, defined in this way, these ROC curves give an idea not only on the ability of the method to separate positives and negatives, but also on its range of applicability and coverage: longer curves represent methods that can be applied to (can generate predictions for) a large number of pairs, and the other way around. So, the ROC curves are generated by cutting the sorted list of scores at different thresholds and plotting the resulting TPR's against FPR's calculated as

$$TPR = Tp/P = sensitivity$$

$$FPR = Fp/N = 1 - specificity$$

where *Tp*, *Fp* and *Tn* are the true positives, false positives and true negatives obtained at a given threshold,

and P and N the total number of positive and negative pairs for that interaction dataset (irrespective of whether the method could be applied for them with that particular set of organisms or not). Note that these parameters can also be interpreted in terms of "sensitivity" and "specificity" as indicated in the formula above.

Additionally, we also evaluated the results in terms of "precision", "recall" (see Results).

Results

As discussed in detail in Methods, for each combination method/interactions-dataset/subset-of-organisms we obtain a ROC plot which represents the capacity of that method for discriminating interacting from non-interacting pairs of proteins (according with the dataset) when using the phylogenetic trees based on that subset of organisms. Figure 2 shows these ROC plots classified by interaction dataset and method. The different curves within each of these plots correspond to the results obtained with the different organism subsets. For "context-mirror" (CM) we show only the results for level 10, which was shown to represent a good threshold of co-evolutionary specificity for predictive purposes [13]. While the results for other levels vary slightly in terms of accuracy/coverage, their behavior respect to the sets of organisms is virtually identical to those of level 10, and hence they are not included here for the sake of clarity.

Each plot in this figure has its own scale to facilitate the comparison between organism sets, which is the final goal of this work. The same figure with all plots in the same scale, which facilitates the comparison between methods, is available as "Additional file 2". The same results represented in terms of F-measure *vs*. score are available as "Additional file 3", together with an explanation of these parameters. Finally, to have a single numerical estimator of the performance of a given method using the trees derived from a given set of

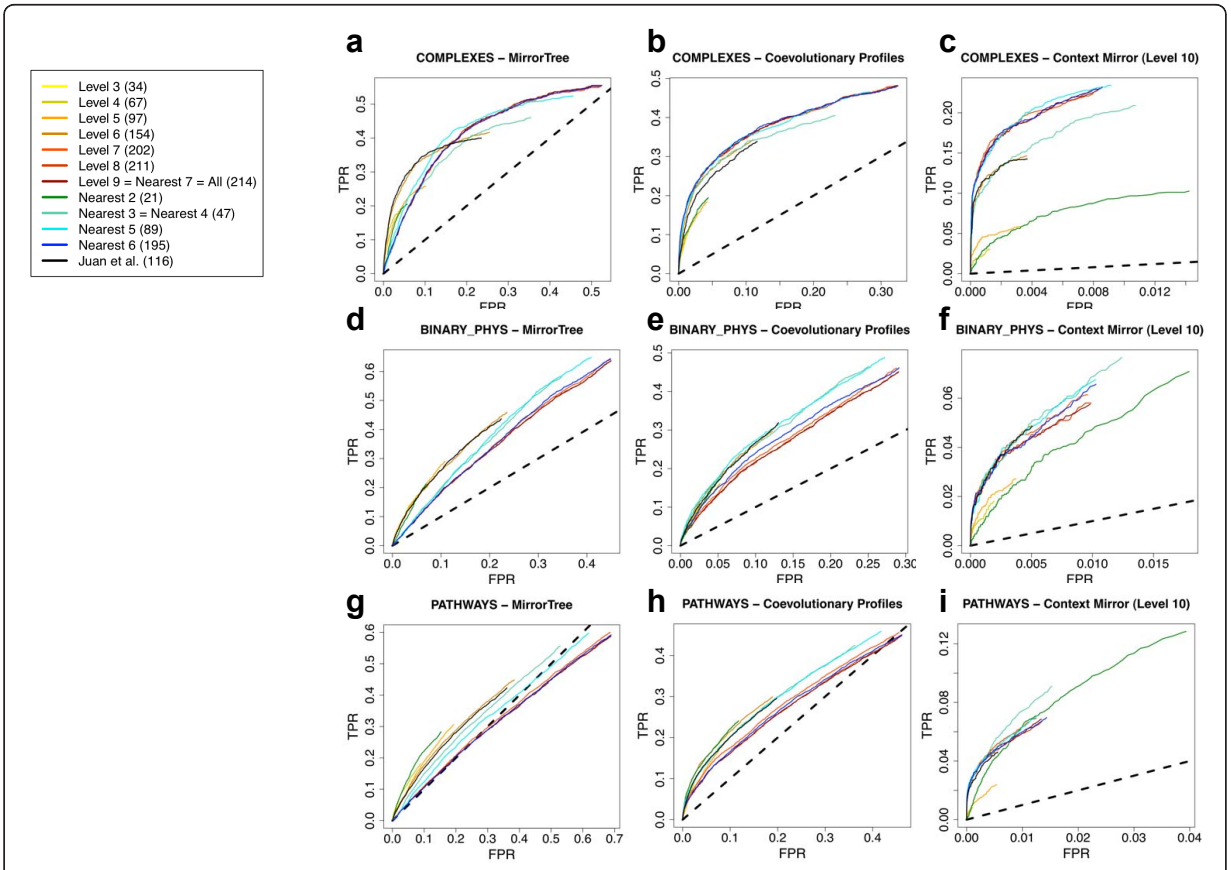


Figure 2 Matrix of partial ROC curves. The partial ROC curves evaluate the performance of a given methodology for a given set of interactions using a given set of reference genomes. The rows represent the interaction datasets and the columns the methods. For a given combination method-dataset, the colored curves represent the organism sets according with the included legend. In the legend, the number of organisms within each subset is indicated within brackets. The dotted line highlights the diagonal of the ROC plots, which represents the background performance expected for a random method.

Herman et al. BMC Bioinformatics 2011, 12:363
http://www.biomedcentral.com/1471-2105/12/363

organisms, the maximum F-measure is shown in the "Additional file 4".

In the "Conclusions" section, we derive some recipes for the future use of these methodologies based on the results shown here.

The most obvious observation is that all these co-evolution based methodologies are able to detect a significant number of interactions of different nature across a wide range of organism sets. This is in line with the growing evidence on the relationship between protein interactions and co-evolution, reported by many groups using diverse datasets and variations of the methodology. Another evident observation is that results can vary largely depending on the set of organisms used for building the trees.

PC is the most stable methodology in the sense that its results are those with the lowest dependence on the organism set, as reflected in the highest similarity between ROC curves (except in extreme cases with few organisms) (Figures 2b, e and 2h). It consistently renders good predictions with the highest independence on the organism set. This could be due to the fact that PC is able to filter artefactual tree-correlations such as those related to phylogenetic bias. CM is globally the best methodology and it produces the highest accuracies, but at the expense of requiring a large number of organisms: its results drastically drop off as we use datasets with low number of organisms (Figures 2c, f and 2h). This effect can be easily explained by the fact that CM requires a rich network of significant inter-protein correlations in order to derive partial correlations. Decreasing the number of organisms reduces the chances of obtaining correlations for many pairs (due to the requirement of 15 organisms in common and also the correlation P-value cutoff), which makes such network sparser and less usable for CM. As previously reported, MT is the methodology with the worst performance and, moreover, it is severely affected by the phylogenetic redundancy in the organism set (Figures 2a, d and 2f). In general, the three methods benefit from using datasets with a large number of organisms. However, for MT, this benefit reaches a point where it enters into conflict with the redundancy issue discussed above resulting in "level6" (≈"Juan et al") being the optimal set. PC and CM implicitly correct phylogenetic redundancy and hence they are more benefited when from using more organisms ("nearest6", "nearest7 = level9 = All").

Another global result is that all methods predict better interactions representing co-presence in macromolecular complexes, followed by binary physical interactions, and being co-presence in metabolic pathways the relationships hardest to detect for all (Figure 2). Within this general trend, each type of interaction seems to be better predicted by a certain set of organisms. In general, complexes are better predicted with datasets including phylogenetically distant organisms, while binary interactions and pathways are better predicted with datasets excluding distant organisms: e.g. follow "nearest5", "6" and "7" in Figure 2.

Examples

We include some examples to illustrate this last result: how using close/distant organisms can drastically affect the predictions. Table 1 contains examples of interactions extracted from the "BINARY_PHYS" dataset which probably correspond to "recent" interactions, as well as others extracted from "COMPLEXES" which probably are "old". The "new" interactions include physical interactions between metabolic enzymes and the interaction between two proteins involved in the division machinery: MinE-MinD [23]. The "old" interactions include some involved in the translation/transcription machinery as well as interactions between ABC transporters. ABC transporters are known to be very ancient systems [24]. The table contains the results that would have been obtained applying the PC method to these cases using as reference sets of organisms "level9" (= all) and "nearest2" (enterobacteriaceae). The correlation coefficient of the PC method indicates the similarity of the co-evolutionary profiles and hence can be seen as a measure of co-evolution. As a measure of performance in detecting the right interactor(s), the "area under the ROC curve" (AUC) [22] is shown. The higher this parameter, the higher are the right interactors (positives) in the sorted list of scores (correlation values). The size of the lists of scores and the number of positives are also indicated. For simplicity and to facilitate the comparison of AUC values, ROC curves are generated here for the positives/negatives which are in the lists, and not taking into account the total number of positives and negatives (as previously done for the ROC curves of Figure 2). It can be seen that "recent" interactions have higher co-evolutionary scores using the "nearest2" dataset than with "level9", and so are the respective predictive performances (AUC). Exactly the opposite happens for the "ancient" interactions: higher co-evolutionary scores and performances are associated to the "level9" set. We follow in detail one of the examples to better understand this table: DPO3A_ECOLI (α subunit of DNA polymerase III) has one reported interaction in the COMPLEXES dataset (with DPO3E_ECOLI, the ε subunit). With the trees constructed based on the "level9" set of organisms, it was possible to apply the PC method to 306 pairs of proteins involving the α subunit (taking into account the requirements and cutoffs described in Methods) one of which is the α-ε pair. Using the "nearest2" set of organisms, it was possible to apply PC to 128 pairs involving DNA pol III α. The co-evolutionary

Table 1 Examples of potentially "new" and "old" interacting pairs of proteins whose co-evolution was evaluated using two sets of organisms

	Protein		Level9(= all)			Nearest2		
			Tot/+	AUC	Interactor (corr)	Tot/+	AUC	Interactor (corr)
"recent" (BINARY_PHYS)	MINE_ECOLI	Cell division topological specificity factor	846/1	0.12	MIND_ECOLI (0.52)	223/1	**0.83**	**MIND_ECOLI (0.60)**
	PABA_ECOLI	Para-aminobenzoate synthase glutamine amidotransferase component II	671/1	0.28	PABB_ECOLI (0.49)	106/1	**0.96**	**PABB_ECOLI (0.96)**
	DHAS_ECOLI	Aspartate-semialdehyde dehydrogenase	760/1	0.17	DNAK_ECOLI (0.48)	384/1	**0.81**	**DNAK_ECOLI (0.90)**
	GSHB_ECOLI	Glutathione synthetase	755/1	0.30	AMPM_ECOLI (0.61)	375/1	**0.93**	**AMPM_ECOLI (0.95)**
"old" (COMPLEXES)	DPO3A_ECOLI	DNA polymerase III subunit alpha	306/1	**0.70**	**DPO3E_ECOLI (0.73)**	128/1	0.11	DPO3E_ECOLI (0.57)
	DPO3E_ECOLI	DNA polymerase III subunit epsilon	357/1	**0.64**	**DPO3A_ECOLI (0.73)**	123/1	0.22	(0.57) *max*
	RPOB_ECOLI	DNA-directed RNA polymerase subunit beta	280/7	**0.82**	**(0.98) *max***	126/4	0.48	(0.93) *max*
	RPOA_ECOLI	DNA-directed RNA polymerase subunit alpha	258/6	**0.81**	**(0.80) *max***	90/3	0.48	(0.93) *max*
	ZNUB_ECOLI	High-affinity zinc uptake system membrane protein znuB	370/2	**1.00**	**(0.87) *max***	129/1	0.36	ZNUC_ECOLI (0.74)
	ZNUC_ECOLI	Zinc import ATP-binding protein ZnuC	386/2	**0.99**	**(0.87) *max***	123/2	0.41	(0.74) *max*
	ZNUA_ECOLI	High-affinity zinc uptake system protein znuA	395/2	**0.98**	**(0.87) *max***	39/1	0.79	ZNUC_ECOLI (0.74)

The co-evolution between these proteins was evaluated using the "level9" and "nearest2" sets of organisms. The total number of pairs involving each protein for which it was possible to make calculations, as well as the number of positives (+) are indicated. The co-evolutionary score with the interactor is also shown (corr). For the cases for which the list contain more than one positive the score is the highest one (max). Finally, the AUC value for the list of scores is also included.

score for α-ε is 0.73 when using the "level9" set of organisms, while it drops to 0.57 when based on "nearest2". As a consequence, there is a much higher proportion of false positives in the sorted list of pairs for "nearest2" compared to "level9" (AUC of 0.11 vs. 0.72). The behavior for the "newer" interactions (e.g. interactions between metabolic enzymes) is exactly the opposite.

Discussion

Our results show that considerable differences in performance are obtained with mirrortree-based methodologies depending on the set of organisms used for building the trees. They also show that it is not always better to use as many genomes as available, as previously assumed. Most of these results have plausible explanations taking into account the type of interaction and the taxonomic distribution of the organisms.

Although the goal of this work is not to compare methods, but organism sets, our results on the performance of the different *mirrortree* variants are in agreement with previous studies [13]. The lower performance of the baseline MT method compared with PC and CM had been already reported and is related to the fact that these two improved methodologies are able to use the information of genome-wide co-evolutionary networks to better detect real co-evolutions as well as implicitly correct phylogenetic biases [13].

The fact that, in general, all methods work better as more genomes are used is not surprising as more co-evolutive information is available for them. Nevertheless, it is important to take into account the issues related to phylogenetic distances and redundancy commented below. PC and CM to some extent correct tree similarities artificially increased by the introduction of redundant genomes (strains, etc.) [13]. That is not the case for *MT* and hence this methodology is especially sensible to this and other phylogenetic biases, some of which can be corrected explicitly [10,11]. The corrections of all these phylogenetic biases implicit in PC and CM make them to be consistently benefited from using more organisms.

The fact that all methodologies render better results for permanent interactions (macromolecular complexes) had been already reported [13]. Actually, for MT and PC, the results for the binary and pathways datasets, in spite of being clearly significant and different from random, might not be of practical applicability in certain prediction scenarios (i.e. if a high precision is required). The explanation for the better predictions of complexes

Herman et al. BMC Bioinformatics 2011, 12:363
http://www.biomedcentral.com/1471-2105/12/363

could be that the evolutionary pressure for co-evolving is expected to be higher in proteins forced to interact permanently than in those with occasional associations. According to these observations macromolecular complexes seem to act as "co-evolutionary units" [13].

Another feature of these macromolecular complexes is that, in general, they represent ancient interactions, compared to transient interactions and functional associations. For this reason, the interaction is expected to occur for all orthologs (interlogs), and hence its associated co-evolutive landmark to be spread through the whole taxonomy. That would explain the observation that better results are obtained for this kind of interactions when including distant organisms within the datasets.

Functional associations and transient interactions are intuitively less prone to yield strong co-evolutions, what would explain the globally lower performances associated to them. Another characteristic of these associations is that, in general, they are "newer" than the macromolecular complexes. It is known that "rewiring" transient interactions is easy and relatively fast in evolutionary terms [25]. For this reason, it may happen that the orthologs of two proteins participating in a transient interaction in a given organism are not interacting in a relatively distant one (they are not true "interlogs") [26,27]. If that is the case, including these "orthologs", which are not interacting and hence not subject to co-evolution, would "dilute" the co-evolutionary signal. This would explain the fact that, for these types of interactions and associations, better results are obtained when using only close organisms, since the interaction is expected to be conserved on them, while it might be absent in taxonomically distant organisms. In other words, many of the *E coli* pathways and transient interactions we are evaluating might be new and hence specific for this microorganism and its close neighbors, and hence the eventual co-evolutions associated to them would be apparent only in these particular genomes. Interestingly, a similar relationship between the "age" of the interactions, their conservation across the taxonomy, and the resulting optimal set of organisms has been reported for the "phylogenetic profiling" method [15].

In some cases it is difficult to disentangle the factors contributing to a given result, for example number of organisms vs. taxonomic criteria used for selecting them. Moreover, it is difficult to quantify and numerically assess the differences of the ROC curves we are using for evaluating performances. For that reason, these curves are evaluated qualitatively and the conclusions presented are based on general trends observed for many curves, instead of particular cases.

A future study aimed at obtaining more insight into the relationship between organism sets and performance should include samplings according with other taxonomic criteria (as well as combinations of them: i.e. combining "nearest"+"level"), and a detailed study of the particular interactions detected and not detected in each experiment (their functional classes, etc).

In the next section, we propose some recipes for the users of these methodologies derived from these results. We plan to implement some of the recipes obtained for the MT method in its recently developed web server [9].

Conclusion

The number of available genomes continues to grow. And the more we know on protein interactions the more we realize that it is a very complex phenomenon with different types of interactions having different characteristics. For these reasons it is increasingly important to "tune" protein interaction prediction methodologies adapting them for each specific application, instead of using the same protocols and data sources in every situation. Many methods and concepts are being built around the reported relationship between similarity of evolutionary histories (co-evolution) and protein interactions. For this reason it is timely to get insight into the different factors affecting such relationship. Among these factors, a critical one not explored previously is the effect of the organism set used to build the trees on the behavior of these methodologies.

Our results allow us to propose a set of simple and general "recipes" for users on which set of organisms to use depending on the type of interactions they want to predict and the genomic information available.

If phylogenetic trees for the whole genome of interest can be calculated (or are already available in some database/resource), use PC and CM instead of MT. If MT has to be used (i.e. trees not available for all the proteins within a genome, lack of computational resources, etc.) the set of organisms to use should be filtered by phylogenetic redundancy. Filtering at the strain or species level seems to be enough.

PC is a sort of "all-road" method since it shows the lowest dependence on the organism set. It is the best option for general situations, when we are not sure which set of organism to use. It is also better than CM in terms of coverage and hence it is more adequate if we are interested in retrieving many interactions at the expenses of bearing more false positives. Moreover, it is computationally less intensive than CM.

CM should be the chosen option when a lot of genomes (as well as enough computational resources) are available and we are interested in detecting a small number of interactions but highly reliable. Not only it renders the best accuracy but additional information on the structure of the co-evolutionary network which offers some clues about the substructure and

functioning of macromolecular complexes is obtained as well [13]. It has to be taken into account that its performance drops drastically when few organisms are available.

Apart from that, if possible it is important to include or exclude distant organisms depending on the type of interactions we try to detect. I.e. to remove phylogenetically distant organisms if we suspect the interactions are not conserved on them ("newer" interactions).

Additional material

Additional file 1: List of organisms in the different subsets and representations of their taxonomic distributions.

Additional file 2: Version of the Figure 2 with all plots in the same scale.

Additional file 3: Results of Figure 2 given in terms of F-measure (the harmonic mean between "precision" and "recall").

Additional file 4: Maximum of the F-measure curves of "Additional file 3". This parameter can be regarded as a single numerical estimator of the performance of a given method/organism set, although it does not encompass all the information of a ROC curve.

Acknowledgements
We sincerely thank the members of the Computational Systems Biology Group (CNB-CSIC) and Profs. Victor de Lorenzo and Miguel Vicente (CNB-CSIC) for interesting discussions. This work was partially funded by projects BIO2009-11966 and BIO2010-22109 from the Spanish Ministry for Science and Innovation, and project KBBE-2007-3-2-08 from the 7FP of the EU. DH was recipient of a "Leonardo Da Vinci" fellowship from the EU. DO is recipient a fellowship from the Basque Country.

Author details
[1]Computational Systems Biology Group, National Centre for Biotechnology (CNB-CSIC), C/Darwin, 3, Cantoblanco, 28049 Madrid, Spain. [2]Structural Bioinformatics Group, Spanish National Cancer Research Centre (CNIO), C/Melchor Fernández Almagro 3, 28029 Madrid, Spain. [3]Centre for Systems Biology (CSB), School of Biosciences, University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK.

Authors' contributions
FP conceived the original idea. FP, DJ and AV designed the experiments. DH, DO and DL implemented the idea and carried out the tests. All authors analyzed the results and contributed writing the manuscript. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Received: 11 March 2011 Accepted: 12 September 2011
Published: 12 September 2011

References

1. Shoemaker BA, Panchenko AR: **Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners.** *PLoS Comput Biol* 2007, **3(4)**:e43.
2. Valencia A, Pazos F: **Computational Methods to predict protein interaction partners.** In *Protein-protein interactions and networks*. Edited by: Panchenko AR, Przytycka TM. London: Springer-Verlag; 2008:67-81.
3. Harrington ED, Jensen LJ, Bork P: **Predicting biological networks from genomic data.** *FEBS Lett* 2008, **582(8)**:1251-1258.
4. Pazos F, Valencia A: **Protein co-evolution, co-adaptation and interactions.** *EMBO J* 2008, **27(20)**:2648-2655.
5. Juan D, Pazos F, Valencia A: **Co-evolution and co-adaptation in protein networks.** *FEBS Lett* 2008, **582(8)**:1225-1230.
6. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO: **Assigning protein functions by comparative genome analysis: Protein pylogenetic profiles.** *Proc Natl Acad Sci USA* 1999, **96**:4285-4288.
7. Date SV, Marcotte EM: **Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages.** *Nat Biotechnol* 2003, **21(9)**:1055-1062.
8. Pazos F, Valencia A: **Similarity of phylogenetic trees as indicator of protein-protein interaction.** *Protein Eng* 2001, **14**:609-614.
9. Ochoa D, Pazos F: **Studying the co-evolution of protein families with the Mirrortree web server.** *Bioinformatics* 2010, **26(10)**:1370-1371.
10. Pazos F, Ranea JAG, Juan D, Sternberg MJE: **Assessing Protein Co-evolution in the Context of the Tree of Life Assists in the Prediction of the Interactome.** *J Mol Biol* 2005, **352(4)**:1002-1015.
11. Sato T, Yamanishi Y, Kanehisa M, Toh H: **The inference of protein-protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships.** *Bioinformatics* 2005, **21(17)**:3482-3489.
12. Kann MG, Jothi R, Cherukuri PF, Przytycka TM: **Predicting protein domain interactions from coevolution of conserved regions.** *Proteins* 2007, **67(4)**:811-820.
13. Juan D, Pazos F, Valencia A: **High-confidence prediction of global interactomes based on genome-wide coevolutionary networks.** *Proc Natl Acad Sci USA* 2008, **105(3)**:934-939.
14. Sun J, Xu J, Liu Z, Liu Q, Zhao A, Shi T, Li Y: **Refined phylogenetic profiles method for predicting protein-protein interactions.** *Bioinformatics* 2005, **21(16)**:3409-3415.
15. Jothi R, Przytycka TM, Aravind L: **Discovering functional linkages and uncharacterized cellular pathways using phylogenetic profile comparisons: a comprehensive assessment.** *BMC Bioinformatics* 2007, **8**:173.
16. Kersey P, Bower L, Morris L, Horne A, Petryszak R, Kanz C, Kanapin A, Das U, Michoud K, Phan I, *et al*: **Integr8 and Genome Reviews: integrated views of complete genomes and proteomes.** *Nucl Acids Res* 2005, **33**:D297-D302.
17. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW: **GenBank.** *Nucleic Acids Res* 2009, **37**:D26-31.
18. Keseler IM, Collado-Vides J, Gama-Castro S, Ingraham J, Paley S, Paulsen IT, Peralta-Gil M, Karp PD: **EcoCyc: a comprehensive database resource for Escherichia coli.** *Nucleic Acids Res* 2005, **33**:D334-337.
19. Goll J, Rajagopala SV, Shiau SC, Wu H, Lamb BT, Uetz P: **MPIDB: the microbial protein interaction database.** *Bioinformatics* 2008, **24(15)**:1743-1744.
20. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Res* 2004, **32(5)**:1792-1797.
21. Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD: **Multiple sequence alignment with the Clustal series of programs.** *Nucleic Acids Res* 2003, **31(13)**:3497-3500.
22. Wikipedia: *ROC analysis* [http://en.wikipedia.org/wiki/Receiver_operating_characteristic].
23. Raskin DM, de Boer PA: **MinDE-dependent pole-to-pole oscillation of division inhibitor MinC in Escherichia coli.** *J Bacteriol* 1999, **181(20)**:6419-6424.
24. Wikipedia: *ATP-binding cassette transporter* [http://en.wikipedia.org/wiki/ATP-binding_cassette_transporter].
25. Shou C, Bhardwaj N, Lam HY, Yan KK, Kim PM, Snyder M, Gerstein MB: **Measuring the evolutionary rewiring of biological networks.** *PLoS Comput Biol* 2011, **7(1)**:e1001050.
26. Aloy P, Ceulemans H, Stark A, Russell RB: **The relationship between sequence and interaction divergence in proteins.** *J Mol Biol* 2003, **332(5)**:989-998.
27. Mika S, Rost B: **Protein-protein interactions more conserved within species than across species.** *PLoS Comput Biol* 2006, **2(7)**:e79.

doi:10.1186/1471-2105-12-363
Cite this article as: Herman *et al.*: Selection of organisms for the co-evolution-based study of protein interactions. *BMC Bioinformatics* 2011 **12**:363.

Molecular BioSystems

RSC Publishing

PAPER

View Article Online
View Journal | View Issue

Incorporating information on predicted solvent accessibility to the co-evolution-based study of protein interactions†

Cite this: *Mol. BioSyst.*, 2013, **9**, 70–76

David Ochoa,[a] Ponciano García-Gutiérrez,‡[a] David Juan,[b] Alfonso Valencia[b] and Florencio Pazos*[a]

A widespread family of methods for studying and predicting protein interactions using sequence information is based on co-evolution, quantified as similarity of phylogenetic trees. Part of the co-evolution observed between interacting proteins could be due to co-adaptation caused by inter-protein contacts. In this case, the co-evolution is expected to be more evident when evaluated on the surface of the proteins or the internal layers close to it. In this work we study the effect of incorporating information on predicted solvent accessibility to three methods for predicting protein interactions based on similarity of phylogenetic trees. We evaluate the performance of these methods in predicting different types of protein associations when trees based on positions with different characteristics of predicted accessibility are used as input. We found that predicted accessibility improves the results of two recent versions of the *mirrortree* methodology in predicting direct binary physical interactions, while it neither improves these methods, nor the original *mirrortree* method, in predicting other types of interactions. That improvement comes at no cost in terms of applicability since accessibility can be predicted for any sequence. We also found that predictions of protein–protein interactions are improved when multiple sequence alignments with a richer representation of sequences (including paralogs) are incorporated in the accessibility prediction.

Received 9th August 2012,
Accepted 9th October 2012

DOI: 10.1039/c2mb25325a

www.rsc.org/molecularbiosystems

Introduction

Computational methods for predicting protein interactions and functional relationships complement experimental techniques in deciphering the networks of protein interactions underlying cellular processes. These techniques are not only faster and cheaper but, in certain situations and for certain types of interactions, their levels of accuracy/coverage are comparable to their experimental counterparts.[1] The tendency now is to combine both approaches in order to obtain reliable interactomes.[2,3]

These computational techniques are based on genomic and sequence features intuitively related to interaction (see ref. 4–7 for recent reviews). A widely used computational approach for detecting interacting proteins is based on similarity of phylogenetic trees (co-evolution). It was repeatedly observed that the phylogenetic trees of interacting proteins are more similar than those of non-interacting ones (see ref. 8, 9 and references therein).

This relationship between protein co-evolution (measured as similarity of trees) and interactions is being exploited in many different ways, ranging from the detailed study of particular interacting families which now can be performed with on-line interactive tools,[10] to the prediction of interactomes in a high-throughput way (*e.g.* ref. 11 and 12), to the prediction of the associations between the members of two protein families known to be related (*e.g.* a family of ligands and the corresponding receptors[13,14]).

The underlying cause for this observed relationship between protein co-evolution and interactions is still a matter of certain debate. The possible explanations range from specific co-adaptation between the interacting partners to general global similarities between their evolutionary rates.[8,15,16] The co-adaptative hypothesis proposes that a long process of specific co-adaptation at the residue

[a] *Computational Systems Biology Group, National Centre for Biotechnology (CNB-CSIC), C/Darwin, 3, Cantoblanco, 28049 Madrid, Spain. E-mail: pazos@cnb.csic.es; Fax: +34 91 5854506; Tel: +34 91 5854669*
[b] *Structural Bioinformatics Group, Spanish National Cancer Research Centre (CNIO), C/Melchor Fernández Almagro 3, 28029 Madrid, Spain*

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c2mb25325a

‡ Present address: Chemistry Department, Universidad Autónoma Metropolitana-Iztapalapa, 09340 México D.F., Mexico.

level (in which interaction de-stabilizing changes in one protein are compensated by changes of similar magnitude in the other) would be the responsible for the observed similarity of evolutionary histories. In the other extreme, it is proposed that this observed similarity could be simply due to the similarity between the evolutionary rates of interacting and functionally related proteins. These two possible explanations for the observed relationship between co-evolution and interactions had been already proposed in the first works dealing with this subject.[17] While these two factors could be jointly contributing to the observed co-evolution, it is possibly the similarity of evolutionary rates that having a major effect, since compensatory changes would need to occur in large numbers in order to really affect the phylogenetic trees.[8]

A number of works have tried, more or less directly, to get some insight into the contribution of co-adaptation to the observed co-evolution.[15,18] The simplest way to approach this problem is to evaluate co-evolution using only the regions of the proteins amenable to co-adaptation (compensatory changes), that is, interaction surfaces (interfaces) or the whole surface, depending on the available information. If co-evolution is (mainly) due to the similarity in evolutionary rates, it would be "spread" through the whole sequence of the proteins, while if it were mainly due to compensatory changes it would be more evident in the surface/interface residues. However, not only surface residues can suffer inter-protein compensatory changes, but also those partially buried or even internal ones *via* indirect and allosteric effects. Moreover, the "intersection" between data on protein three-dimensional (3D) structures and interactions is not high, leading to small or eventually biased datasets to perform these studies. The scarcity in 3D data has another effect: if a methodology is eventually developed which combines co-evolution with structural information (solvent accessibility) for improving the accuracy in predicting interactions, its range of applicability would drop drastically compared to its counterparts which require only sequence information. Based on the above, it would be desirable to study the effect of incorporating predicted solvent accessibility information on co-evolution methods, instead of the "real" solvent accessibility extracted from experimental 3D structures. Predicted solvent accessibility can be obtained for any sequence, and with good levels of accuracy: above 75% for two-state predictions ("buried/exposed").[19,20]

In this work we assess for the first time the effect of including predicted solvent accessibility information on the results and range of applicability of three co-evolution based methods for predicting protein interactions. We used a number of datasets representing different types of interactions (physical, functional, ...) as gold standards in order to interpret the results in terms of the type of interaction of interest.

Methods

We aim to evaluate the effect of the incorporation of information on predicted solvent accessibility in the performance of three *mirrortree*-related methods in predicting interactions of different nature. This has been done by generating, for all proteins in the model organism *E. coli*, different sets of phylogenetic trees constructed using (i) the whole protein and (ii) only the protein residues above certain thresholds of predicted accessibility, and evaluating the performance of the methods based on these different trees. In order to evaluate the performance we used different datasets of protein interactions representing interactions of different nature (*e.g.* physical and functional). The process is illustrated in Fig. 1, and details are given below.

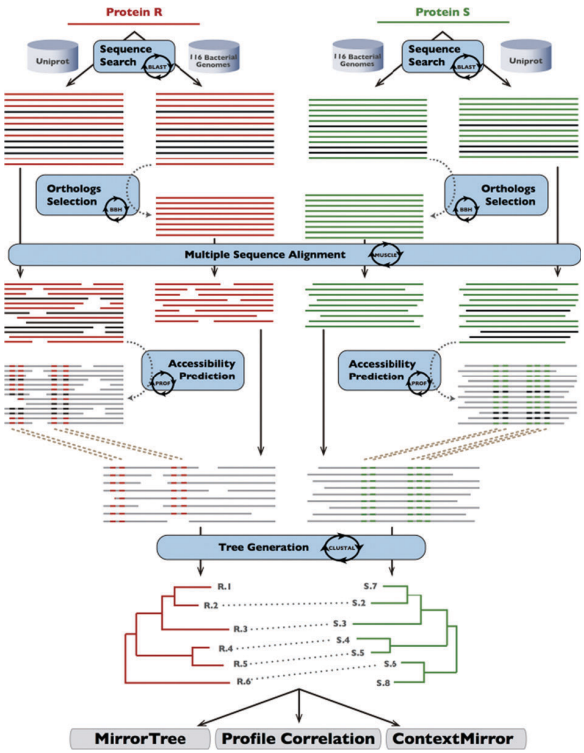


Fig. 1 Scheme of the methodology. In order to evaluate the co-evolution between proteins R and S based on their residues fulfilling a given predicted accessibility criterion, the first step is to look for their orthologs in a set of 116 fully sequenced genomes. For each protein, a multiple sequence alignment is generated with these orthologs, which will serve as a basis for the generation of the trees. In parallel, another multiple alignment is generated for the same protein based on the homologs found in the whole Uniprot database (hence including orthologs and paralogs). This second alignment will be used for the prediction of solvent accessibility. A tree is generated based on the first alignment but using only the positions with a given predicted accessibility criterion. The trees generated in this way are the input for the three methods for evaluating co-evolution.

Solvent accessibility prediction

First, for each *E. coli* protein, a list of candidate homolog protein sequences was retrieved searching with BLAST[21] in the non-redundant Uniprot database.[22] Sequences with an *E*-value greater than 1×10^{-4} or an identity (based on the BLAST alignment) less than 20% were excluded. Alignment coverages lower than 60% (either respect to the hit or the query protein) were also excluded.

A multiple sequence alignment (MSA) for the remaining sequences was generated using MUSCLE.[23] The identity of each aligned sequence with the *E. coli* reference sequence was then calculated using only positions with less than 90% gaps. If this identity was less than 20%, the sequence was discarded. Additionally, sequence redundancy was removed at 95% to avoid the overrepresentation of some sequences, which could influence the accessibility predictions.

Finally, this multiple sequence alignment was used as input for the PROF program for predicting solvent accessibility,[24,25] and the predictions for the columns of the MSA were mapped to the positions of the original *E. coli* protein. For comparative purposes equivalent accessibility predictions were also generated based on the MSAs of orthologs used for constructing the phylogenetic trees (described in the next section).

Generation of phylogenetic trees

We used a set of 116 fully sequenced organisms previously used in other works[11,26] to look for orthologs of *E. coli* proteins and construct the trees based on them. This set does not contain very similar organisms, thus avoiding phylogenetic redundancy.

We used the ''BLAST best bi-directional hit'' criterion for detecting the ortholog of a given *E. coli* protein in each genome, with an *E*-value cut-off of 1×10^{-5}, and requiring an alignment coverage of 70%. All orthologs found for this *E. coli* protein were aligned with MUSCLE[23] using the default parameters of this program. Then, a phylogenetic tree was generated from this alignment using the neighbor-joining algorithm implemented in ClustalW,[27] excluding the gaps for the distance calculation.

Equivalent trees were generated but using only the positions of the alignment fulfilling the following criteria of predicted accessibility:

- eRIA0: positions predicted as accessible by PROF with any value of ''reliability''.
- eRIA3: positions predicted as accessible with reliability ≥ 3 (PROF reliability values range from 0 to 9).
- pACC2, pACC12 and pACC50: positions with a predicted solvent accessible surface ≥ 2, 12 and 50 Å^2, respectively.

Finally, distance matrices containing the pair-wise distances between all orthologs were generated for the original tree (based on the whole length of the protein) as well as for these trees based on (predicted) accessible positions. These distances are calculated by summing the lengths of the branches separating the corresponding leaves. These distance matrices are the input for the *mirrortree*-based methods described in the next point.

Prediction of protein interactions based on phylogenetic trees

The original *mirrortree* (MT) approach[17] evaluates the co-evolution between two protein families by calculating the linear correlation coefficient between the values of their corresponding distance matrices. A minimum of 15 species in common is required in order to evaluate a given pair of proteins. Moreover, only correlation values supported by a tabulated *P*-value of 1×10^{-5} or better are used.

The *profile-correlation* (PC) method[11] takes as input the *mirrortree* raw scores for all pairs of proteins in a given organism. Hence, in this case the input is a squared matrix the size of the *E. coli* proteome with the correlation values for all pairs of proteins (actually, those with 15 or more organisms in common and supported by a *P*-value $\leq 1 \times 10^{-5}$). A row in this matrix, known as ''co-evolutionary profile'', represents the co-evolutionary behaviour of a protein respect to the rest of the proteome. Within the context of the PC method, the co-evolution between two proteins is re-evaluated as the correlation between their corresponding co-evolutionary profiles, with the same significance thresholds used for the original *mirrortree*. The idea is that two proteins whose trees are similar and, additionally, that tend to be similar to the same set of proteins (and dissimilar to the complementary set) are more likely to represent a case of true co-evolution.

The *context-mirror* (CM) method[11] takes into account the influence of ''third proteins'' in a given co-evolutionary signal observed for a given pair of proteins using a partial correlation criterion. In this way it is possible to separate specific co-evolution (particular to a given pair of proteins) from general co-evolutionary trends involving many proteins. For a given pair of proteins, this method produces results at different ''levels'' of specificity, being ''level 1'' the one representing the most specific co-evolution.

Datasets of protein interaction and functional relationship

The performance of the three methods when fed with phylogenetic trees generated with residues of different predicted accessibility was evaluated using three datasets representing protein interactions of different nature in *E. coli* as gold standard.

- Binary physical direct interactions obtained from MPIDB.[28] This database contains binary interactions manually curated from the literature or imported from other databases. We retrieved the 2103 binary interactions between 1538 different *E. coli* proteins stored on it.
- Physical (sometimes indirect) interactions inferred as co-presence in experimentally determined macromolecular complexes obtained from EcoCyc.[29] This dataset contains 1354 experimentally determined interactions between 591 proteins.
- Functional interactions inferred as membership in the same metabolic pathways, also taken from the EcoCyc. This dataset contains 4419 relations between 719 proteins.

In the three cases, the sets of negatives (pairs of proteins regarded as non-interacting) were constructed by generating all possible pairs between the proteins in the corresponding positive (interacting) sets, excluding those pairs already annotated as interacting.

Performance evaluation

For each combination of a method, an input set of trees (generated from residues of different predicted accessibility) and an interaction dataset we obtain a list of protein pairs, sorted by the score of the corresponding method. Each pair can be labelled as positive or negative depending on whether it is a reported interaction in that particular dataset or not. A combination

method-set of trees will be better for predicting interactions (for that particular set of interaction evidences) as the positives tend to cluster at the top of these sorted lists (associated to high scores) and the other way around for the negatives.

The Area Under the ROC Curve (AUC) was calculated for these lists, as a global estimator of the accuracy and coverage of the corresponding predictions. The ROC (''receiver operating characteristic'') analysis[30] generates a plot of ''true positives rate'' (TPR) against ''false positives rate'' (FPR) when varying the classification threshold (score of the method). Curves above the diagonal in this plot represent methods with some discriminative power, being this discriminative capacity better as the curve gets closer to the top-left corner of the plot. Consequently, areas under these curves range from 0.5 (random classifier, diagonal in the plot, positives and negatives uniformly distributed through the list) to 1.0 (perfect classifier, all positives at the top of the list). ROC analysis was performed with the ROCR library of the R statistical package (http://www.r-project.org).

Results and discussion

Fig. 2 and Fig. S1 (ESI†) show the performance (AUC value) of the three *mirrortree*-based methods, when using the phylogenetic trees constructed from residues of different predicted accessibility, and evaluated based on the three different datasets of protein interactions.

As previously seen,[11,26] *mirrotree*-based methods predict better physical interactions (binary and complexes) than functional associations (*e.g.* pathways). Within physical interactions, those representing co-membership to macromolecular complexes are better detected than those representing binary (eventually transient) interactions. About the methods, the PC and CM methods work

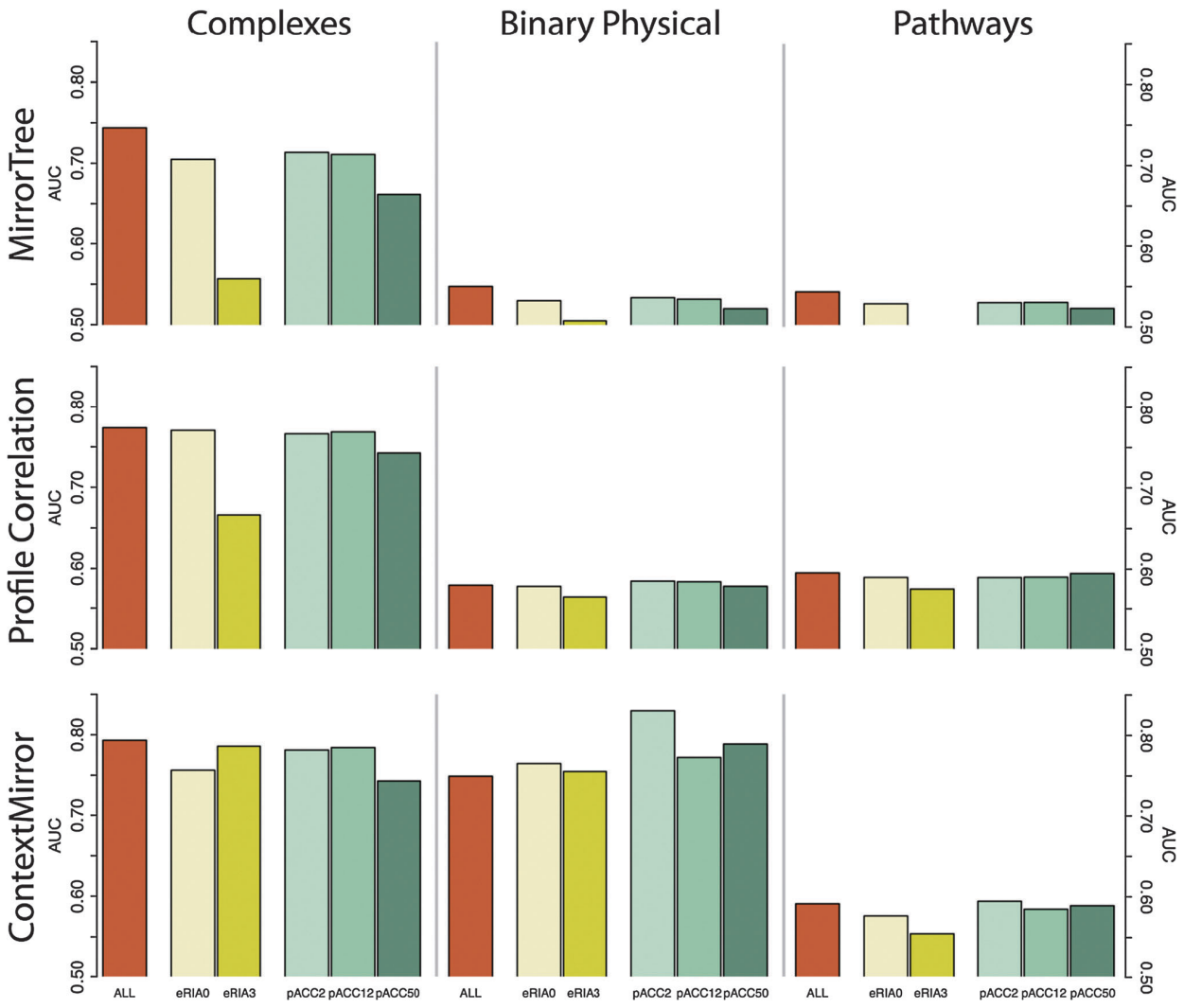


Fig. 2 Performances for different combinations of: phylogenetic tree comparative methods, interaction evidence and predicted accessibility filter. Performance is evaluated as the ''Area Under the [ROC] Curve'' (AUC). The same figure with different scales for each plot is available as Fig. S1 (ESI†). Equivalent figures with the results obtained using predicted accessibility derived from MSAs of orthologs are available as Fig. S2 and S3 (ESI†).

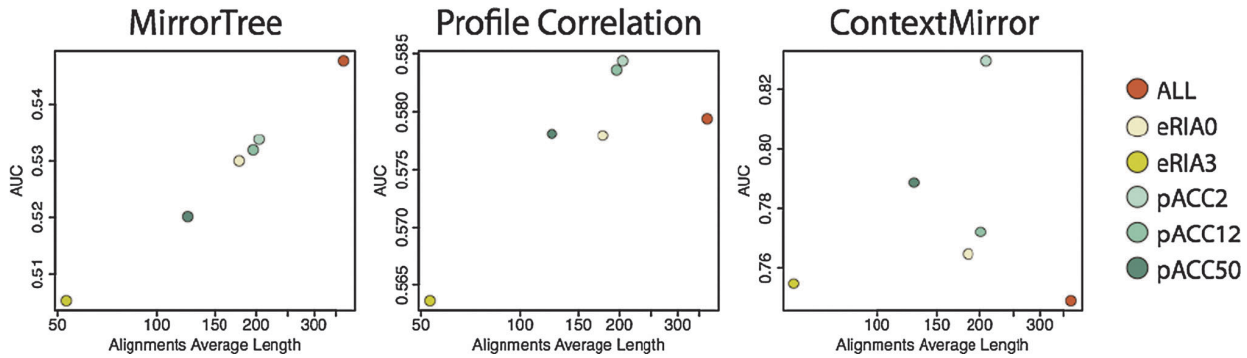


Fig. 3 Relationship between the performances and the lengths of the virtual alignments. The length of the virtual alignment is the number of positions (fulfilling a given predicted accessibility criteria – colors) used for deriving the trees. The data shown here are for binary physical interactions. The corresponding plots for the other interaction datasets are available as Fig. S4 (ESI†).

better than the original MT approach, which presents "usable" levels of performance only for detecting interactions of macromolecular complexes.

Predicted accessibility only helps the PC and CM methods in detecting binary physical interactions

For most cases, the use of predicted solvent accessibility within *mirrortree*-based methodologies worsens the results (Fig. 2 and Fig. S1, ESI†). The AUC values for these methods working with trees derived from different sets of (predicted) accessible residues are worse than those based on full sequences.

Interestingly, for the case of binary physical interactions, the results of the PC and CM methods are improved when using predicted solvent accessibility. The best results are obtained when using all residues with a minimum of solvent accessible area ("pACC2", area $\geq$ 2 Å^2). Restricting to residues predicted to be highly accessible ($\geq$ 12 and $\geq$ 50 Å^2), or those predicted as "accessible" by PROF's two-state predictor (eRIA0 and eRIA9) works worse than with $\geq$ 2 Å^2.

For most cases, there is a correlation between the performances obtained with the trees based on different predicted accessibilities and the average lengths of the virtual alignments used for deriving them (number of positions fulfilling that particular accessibility cut-off) (Fig. 3 and Fig. S4, ESI†). This trend is broken for the results of the PC and CM methods predicting direct physical interactions: in these two cases pACC2 renders the better results in spite of not having the largest virtual alignments (Fig. 3). This general decrease in performance when incorporating predicted accessibility could be partly due to the intrinsic errors associated with the prediction. Nevertheless, it is probably more related to the largest contribution of the similarity of evolutionary rates to the observed co-evolution (see above): the co-evolutionary signal would be spread through the whole sequence and not restricted to certain parts (surfaces, *etc.*) This is reinforced by the observation that, in general, performances correlate positively with the number of positions used for building the trees.

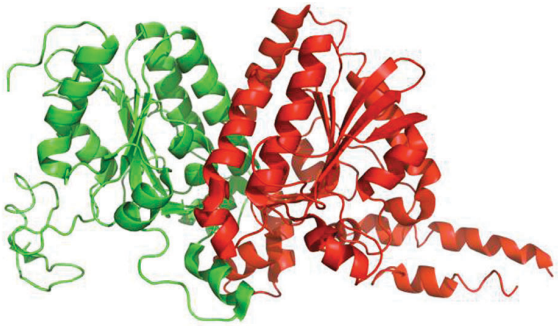
Our interpretation for the fact that accessibility predictions do not help the original MT (but the other way around) is that this methodology is mainly detecting non-specific co-evolution associated with global similarities in evolutionary rates (reflected in the whole sequence, as commented above). The more recent PC and CM methods benefit by the use of predicted accessibility when applied for the prediction of binary direct physical interactions. These two methods have been previously associated with the detection of more specific co-evolution,[11] cases where the co-adaptive part of the co-evolutionary signal is probably higher. In the same line, these specific co-evolutions (with larger proportions of co-adaptations) are intuitively more related to direct physical interactions than, for example, to those relating the members of macromolecular complexes.

It is also interesting that the set of predicted accessible residues which renders the best results include residues with a minimal predicted solvent accessible area (> = 2 Å^2), which are better than those with higher levels of predicted accessibility (> = 12 Å^2, > = 50 Å^2). That could be explained by the fact that co-adaptation is not necessarily restricted to totally exposed residues but can also happen between their neighbours or even buried residues (through allosteric effects).

It is better to use accessibility predicted from MSAs constructed for this purpose, than that based on MSAs with orthologs only

Fig. S2 and S3 (ESI†) show the same AUC results as Fig. 2 and Fig. S1 (ESI†) but using accessibility predicted from the same alignments used for constructing the phylogenetic trees input of the *mirrortree*-based methods (composed by orthologs only). The general drop in performance detected when incorporating accessibility information can be observed to be sharper here. Moreover, for those cases in which predicted accessibility improved the results (PC and CM predicting physical interactions, previous point) the improvement obtained with these alignments of orthologs is still present but smaller. Therefore, accessibility predicted from the same alignments used for constructing the phylogenetic trees renders worse results than that predicted from MSAs constructed *ad hoc* for this purpose.

The fact that accessibility predicted from "richer" alignments (including eukaryotic sequences and eventually paralogs) is



	ALL	eRIA0	eRIA3	pACC2	pACC12	pACC50
MirrorTree	0.9083	0.8982	0.8331	0.8998	0.9069	0.8924
Profile Correlation	0.9516	0.9531	0.9435	0.9605	0.9592	0.9328
ContextMirror	0.6068	0.6650	0.5410	0.6818	0.6828	0.5407

Fig. 4 Example illustrating the effect of incorporating predicted solvent accessibility on the evaluation of tree similarity. The structure of the complex between the α and β chains of *E. coli* acetyl-CoA carboxylase carboxyl transferase is shown in ribbon representation. The table contains the scores of the three methods for this interacting pair of proteins based on the trees derived with the six different criteria of predicted accessibility.

better in helping these co-evolution based methods than that based on alignments containing only bacterial orthologs was expected. It was previously shown that the quality of the MSA is critical for obtaining good sequence-based predictions of protein features such as accessibility or secondary structure.[19] Nevertheless, we wanted to make a test with MSAs of orthologs due to a methodological reason: these MSAs have to be generated in order to apply *mirrortree* and related methods. Consequently, if the accessibility predicted from them turned out to perform similarly to that predicted from richer alignments, it would be trivial to add this accessibility prediction step to current *mirrortree* workflows. Unfortunately, although some improvement is obtained with that accessibility, the best results are obtained when using that predicted from richer alignments. Consequently, in order to obtain these optimal results the workflow has to be "bifurcated", generating one alignment for tree construction and another one for accessibility prediction, as shown in Fig. 1.

Example

For illustrative purposes only, we include an example of an interacting pair of proteins for whose co-evolution is more evident when evaluated using solvent accessible predicted residues. Fig. 4 shows the results of *mirrotree* and related methods evaluating the co-evolution between the α and β subunits of the *E. coli* acetyl-CoA carboxylase carboxyl transferase. It can be seen that the similarity between the evolutionary histories of these two interacting proteins is more evident when evaluated from trees constructed using the residues predicted as accessible, except for the original MT method. For example, the score of the ContextMirror method increases from 0.60, when it is based on the trees derived from the whole sequence of these proteins, to 0.68 (trees based on predicted solvent accessible residues).

Conclusions

The underlying cause for the observed relationship between protein co-evolution and interactions is still not totally clear. The possible explanations range from unspecific co-evolution due to the similarity of evolutionary rates of interacting proteins, to specific co-adaptation involving inter-protein compensatory changes.[8,16] It is possibly the first factor the one playing a major role since evolutionary rate and interactions have been previously related through a number of direct and indirect paths.[15,31] The co-evolution observed in pairs of functionally related proteins which do not necessarily interact physically (*e.g.* ref. 32 and 33) is also easier to understand under this hypothesis. Nevertheless, compensatory changes have been repeatedly observed in protein interfaces (*e.g.* see ref. 8) and are surely playing a role in the co-evolution of interacting proteins at a local level. However, it is difficult to conceive these changes as mostly responsible for the observed tree similarity, since a very large number of such compensatory changes would be necessary to have an effect on the shapes of the trees. Previous studies tried to disentangle these two factors by comparing the co-evolution of protein regions amenable to compensatory changes (surfaces and interfaces) to that of the whole protein length.[15,18] In this work we tackle this problem but using predicted solvent accessibility, instead of real surfaces.

We have demonstrated that using predicted solvent accessibility helps in the co-evolution based prediction of protein interaction under some circumstances. Besides the implications of these results for the debate on the contribution of co-adaptation to the observed relationship between tree similarity and interactions, this work has also practical implications for the application of these methodologies, and these are not only related to the improvement in the prediction of protein interaction. Since this method goes on a step further in the detection of the protein regions actually co-evolving, it opens interesting possibilities for studying how the residues at the interfaces change and co-adapt during evolution. This could give some insight into the physico-chemical basis of protein interactions since the coordinated changes at the interfaces would provide a picture of possible interactions modes for a particular protein family. Moreover co-evolution has been proposed as a mechanism for maintaining interactions between proteins while allowing them to change at the same time. In many interacting protein families co-evolution is reflected in a set of specific surface residues which concomitantly change in both interacting partners. These residues are good candidates for mutagenesis experiments aimed at switching the interaction specificity of the proteins and/or adapting them to new interaction partners.

It is also important to highlight that the improvement obtained when incorporating predicted solvent accessibility does not have any associated cost in terms of coverage/applicability, since accessibility predictions can be generated for any sequence.

Acknowledgements

We want to thank Prof. Burkhard Rost and his group (currently at Technische Universität München) for help with local installations of PROF. We also want to thank the members of the

Computational Systems Biology Group (CNB-CSIC) for insightful discussions. This work was funded by a project from the Spanish Ministry for Economy and Competitiveness (BIO2010-22109). DO is recipient a fellowship from the Basque Country. PGG also thanks Mexican CONACyT for a graduate studies grant.

References

1 C. von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields and P. Bork, *Nature*, 2002, **417**, 399–403.
2 I. Lee, S. V. Date, A. T. Adai and E. M. Marcotte, *Science*, 2004, **306**, 1555–1558.
3 C. von Mering, M. Huynen, D. Jaeggi, S. Schmidt, P. Bork and B. Snel, *Nucleic Acids Res.*, 2003, **31**, 258–261.
4 B. A. Shoemaker and A. R. Panchenko, *PLoS Comput. Biol.*, 2007, **3**, e43.
5 A. Valencia and F. Pazos, in *Protein–protein interactions and networks*, ed. A. R. Panchenko and T. M. Przytycka, Springer-Verlag, London, 2008, pp. 67–81.
6 E. D. Harrington, L. J. Jensen and P. Bork, *FEBS Lett.*, 2008, **582**, 1251–1258.
7 M. N. Wass, A. David and M. J. Sternberg, *Curr. Opin. Struct. Biol.*, 2011, **21**, 382–390.
8 F. Pazos and A. Valencia, *EMBO J.*, 2008, **27**, 2648–2655.
9 D. Juan, F. Pazos and A. Valencia, *FEBS Lett.*, 2008, **582**, 1225–1230.
10 D. Ochoa and F. Pazos, *Bioinformatics*, 2010, **26**, 1370–1371.
11 D. Juan, F. Pazos and A. Valencia, *Proc. Natl. Acad. Sci. U. S. A.*, 2008, **105**, 934–939.
12 E. R. Tillier and R. L. Charlebois, *Genome Res.*, 2009, **19**, 1861–1871.
13 A. K. Ramani and E. M. Marcotte, *J. Mol. Biol.*, 2003, **327**, 273–284.
14 J. M. Izarzugaza, D. Juan, C. Pons, J. A. Ranea, A. Valencia and F. Pazos, *Nucleic Acids Res.*, 2006, **34**, W315–W319.
15 L. Hakes, S. Lovell, S. G. Oliver and D. L. Robertson, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**, 7999–8004.
16 S. C. Lovell and D. L. Robertson, *Mol. Biol. Evol.*, 2010, **27**, 2567–2575.
17 F. Pazos and A. Valencia, *Protein Eng.*, 2001, **14**, 609–614.
18 M. G. Kann, B. A. Shoemaker, A. R. Panchenko and T. M. Przytycka, *J. Mol. Biol.*, 2009, **385**, 91–98.
19 B. Rost, in *Structural Bioinformatics*, ed. P. E. Bourne and J. Gu, Wiley-Blackwell, 2nd edn, 2009, pp. 679–714.
20 I. Y. Y. Koh, V. A. Eyrich, M. A. Marti-Renom, D. Przybylski, M. S. Madhusudhan, N. Eswar, O. Grana, F. Pazos, A. Valencia, A. Sali and B. Rost, *Nucleic Acids Res.*, 2003, **31**, 3311–3315.
21 S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman, *Nucleic Acids Res.*, 1997, **25**, 3389–3402.
22 U. Consortium, *Nucleic Acids Res.*, 2009, **37**, D169–D174.
23 R. C. Edgar, *Nucleic Acids Res.*, 2004, **32**, 1792–1797.
24 B. Rost, *Methods Enzymol.*, 1996, **266**, 525–539.
25 B. Rost, in *The Proteomics Protocols Handbook*, ed. J. E. Walker, Humana, Totowa, NJ, 2005, pp. 875–901.
26 D. Herman, D. Ochoa, D. Juan, D. Lopez, A. Valencia and F. Pazos, *BMC Bioinf.*, 2011, **12**, 363.
27 R. Chenna, H. Sugawara, T. Koike, R. Lopez, T. J. Gibson, D. G. Higgins and J. D. Thompson, *Nucleic Acids Res.*, 2003, **31**, 3497–3500.
28 J. Goll, S. V. Rajagopala, S. C. Shiau, H. Wu, B. T. Lamb and P. Uetz, *Bioinformatics*, 2008, **24**, 1743–1744.
29 I. M. Keseler, J. Collado-Vides, S. Gama-Castro, J. Ingraham, S. Paley, I. T. Paulsen, M. Peralta-Gil and P. D. Karp, *Nucleic Acids Res.*, 2005, **33**, D334–D337.
30 T. Fawcett, *Pattern Recogn. Lett.*, 2006, **27**, 861–874.
31 H. B. Fraser, A. E. Hirsh, L. M. Steinmetz, C. Scharfe and M. W. Feldman, *Science*, 2002, **296**, 750–752.
32 Z. Liang, M. Xu, M. Teng, L. Niu and J. Wu, *FEBS Lett.*, 2010, **584**, 4237–4240.
33 N. L. Clark, E. Alani and C. F. Aquadro, *Genome Res.*, 2012, **22**, 714–720.

Acknowledgements

Llegado el momento de agradecer todos los que han hecho que esta tesis se haga realidad me vienen a la mente muchas personas. Seguramente no me acordaré de todos y sin duda, en unas pocas palabras no les voy a conceder todo el rédito que en realidad se merecen. De algunos he aprendido mucho de otros muchísimo y todos en un mayor o menor parte han hecho que este trabajo salga adelante. Seguramente es debido a todos los que aquí menciono que haya disfrutado haciendo este trabajo.

En primer lugar y sobre todo agradecer a Sito por todo su esfuerzo y confianza. Han sido muchos años de trabajo en los que ni en un solo día he tenido la menor queja. Sin duda has hecho fácil toda mi andadura en el grupo desde el día en que llegué y probablemente es algo que se valore más retrospectivamente. He aprendido mucho de tí y siempre has estado ahí para apoyarme en lo que fuera necesario y para corregirme en lo que me equivocaba. Gracias por todo.

En parte este agradecimiento se extiende a los que han estado conmigo a diario desde el día en que aterricé en el grupo. Dani gracias por tu ayuda en la representación de los árboles taxonómicos. Gracias por todos las discusiones de asuntos de los que no teníamos ni idea, por tus eternas lecciones sobre la programación orientada a objetos, y por la diplomacia entre muchas otras cosas. JC gracias por enseñarme a ver el mundo de otra manera y por los millones de favores que siempre me has hecho no importara cuánto costara o qué hora fuera. Mucha suerte. Mónica siempre me quedaré con las ganas de haber trabajado más juntos. Muchas gracias por todos tus consejos. Gracias a Natalia por hacer más felices nuestras interminables sesiones a deshoras. Oli y David gracias por las interminables discusiones de lo más peregrinas durante los cafés. Ha sido un placer. Gracias Toño por esos eternos debates científicos o filosóficos que siempre teníamos. Gracias Jose Manuel por los consejos que siempre me has dado. Gracias a Trivi, Luis, Ponciano, Aldo, Pablo por todo lo que pude aprender de vosotros.

Mucha parte de culpa de esta tesis la tienen Alfonso y David. Gracias David por todo lo que he aprendido de ti que ha sido mucho. Ha sido un placer compartir contigo las interminables charlas en las que acababamos con más dudas que respuestas. Espero y confío en que vengan muchas más en el futuro. Lo dicho, un placer.

Thank you Dorota for your contribution to the organisms selection project selection. It's been a while but I really enjoyed working with you. Good luck!

Trey, many thanks for hosting me in San Diego. Thank you Janusz and Emre for all I've learned from you on these years. It's been a puzzling and stimulating work but I really enjoyed it. Many thanks as well as to the rest of laboratory at UCSD who made my stay in SD so great.

Muchas gracias también al programa de formación de personal investigador del Gobierno Vasco por hacer que esta aventura haya sido posible. A las distintas instituciones que han financiado parte de nuestra investigación. Al CSIC, al personal del CNB, muchas gracias por contribuir a este trabajo.

Muchas gracias a aita y ama. Porque siempre me han apoyada hiciese lo que hiciese. A Aida, Lara y Charly porque siempre están ahí y no hace falta llamarles. A mis amigos. Gracias Ferni y Alberto por pasar algunos de estos años juntos y por acogerme en mis periodos de sin techo. Lo mismo Jon, Mati y Murga. Finalmente y no por eso menos importante, gracias Nere por aguantarme en estos últimos tiempos. Por tu inestimable ayuda y por motivarme hasta en los peores momentos. Gracias sobre todo por compartir conmigo los retos futuros.