Critical Review

Protein Coadaptation and the Design of Novel Approaches to Identify Protein–Protein Interactions

Mario A. Fares[1,2], Mario X. Ruiz-González[1] and Juan Pablo Labrador[2,3]

[1] *Department of Abiotic Stress, Group of Integrative and Systems Biology, Instituto de Biología Molecular y Celular de Plantas (CSIC-Universidad Politécnica de Valencia), Valencia, Spain*
[2] *Department of Genetics, Trinity College, University of Dublin, Dublin 2, Dublin, Ireland*
[3] *Institute of Neuroscience, Trinity College, University of Dublin, Dublin 2, Dublin, Ireland*

Summary

Proteins rarely function in isolation but they form part of complex networks of interactions with other proteins within or among cells. The importance of a particular protein for cell viability is directly dependent upon the number of interactions where it participates and the function it performs: the larger the number of interactions of a protein the greater its functional importance is for the cell. With the advent of genome sequencing and "omics" technologies it became feasible conducting large-scale searches for protein interacting partners. Unfortunately, the accuracy of such analyses has been underwhelming owing to methodological limitations and to the inherent complexity of protein interactions. In addition to these experimental approaches, many computational methods have been developed to identify protein–protein interactions by assuming that interacting proteins coevolve resulting from the coadaptation dynamics between the amino acids of their interacting faces. We review the main technological advances made in the field of interactomics and discuss the feasibility of computational methods to identify protein–protein interactions based on the estimation of coevolution. As proof-of-concept, we present a classical case study: the interactions of cell surface proteins (receptors) and their ligands. Finally, we take this discussion one step forward to include interactions between organisms and species to understand the generation of biological complexity. Development of technologies for accurate detection of protein–protein interactions may shed light on processes that go from the fine-tuning of pathways and metabolic networks to the emergence of biological complexity.

IUBMB *Life*, 63(4): 264–271, 2011

Received 17 December 2010; accepted 22 February 2011

Address correspondence to: Dr. Mario A. Fares, Department of Abiotic Stress, Group of Integrative and Systems Biology, Instituto de Biología Molecular y Celular de Plantas (CSIC-Universidad Politécnica de Valencia), Valencia, Spain. E-mail: mfares@ibmcp.upv.es (or) Dr. Juan Pablo Labrador, Department of Genetics, Trinity College, University of Dublin, Dublin 2, Dublin, Ireland. E-mail: jp.labrador@tcd.ie.

ISSN 1521-6543 print/ISSN 1521-6551 online
DOI: 10.1002/iub.455

Keywords protein interaction; coevolution; yeast 2 hybrid; coadaptation.

INTRODUCTION

A formidable challenge in the proteomics and systems biology era is to understand how the genetic information links gene products to functions that are essential for the viability and communication of cells with the environment. Genes codify proteins (functions) that generally interact in a fine-tuned way with other proteins in the cell or between cells to perform a particular task (*1*). They rarely perform their function independently but they form part of complex networks of interaction. The order and timing at which these interactions occur is fundamental in triggering the biochemical reactions of the cell; otherwise, its violation leads to the emergence of aberrant phenotypes. Moreover, these interactions require overcoming numerous error-testing steps that go from the synthesis of proteins and their post-translational modification to the formation of protein complexes (*2*).

The type of interacting partners for a protein determines its impact on cell viability because highly connected proteins are expected to participate in more functions than lowly connected ones. Therefore, finding partners for a particular protein may aid to determine protein's function, calculate the consequences of knocking this protein down or identify candidate disease genes for novel therapies. The identification, description and understanding of protein interactions is, therefore, the cornerstone of fundamental cell biology and medicine.

Finding protein complexes has been traditionally conducted using affinity purification methods (*3*). These methods utilize tagged baits that make possible retrieving protein complexes and identifying them using other approaches such as mass spectrometry (*4, 5*). In combination with these methods, the yeast two-hybrid system (Y2H) (*6, 7*) and protein chips (*8*) have been extensively utilized to identify and characterize protein interactions. The Y2H strategy starts coupling two proteins named the bait and prey, to two halves of a transcription factor and expressed in yeast. The transcription factor can only activate a reporter gene when both of the proteins, prey and bait, interact reconstituting the DNA binding domain and the transactivation domain of the transcription factor.

Many groups have conducted high-throughput screenings of protein–protein interactions. For example, Gavin and colleagues (*9*) performed a genome-wide screening of protein complexes in the budding yeast *Saccharomyces cerevisiae* using affinity purification and mass spectrometry. Their approach provided a *de novo* characterization of 257 novel protein complexes, 73% out of which were known. Significantly, they found no evidence for 74 of the known protein complexes, raising concern on the validity of many of the complexes that were previously assumed to exist (*9*). In a later study (*10*), authors implemented a more exhaustive analysis of the interactome (the set of protein–protein interactions that occurs in a cell at any given time) in *S. cerevisiae* in which they processed 4,562 different tagged proteins using tandem affinity purification. They identified 7,123 protein interactions that involve 2,708 proteins. Over 270 of those complexes were not previously detected and 429 additional interactions between complexes were reported.

Several attempts to identify and discard false positives have shown that the accuracy of the experimental approaches to identify binary protein interactions is underwhelming. For example, using a method designed to computationally assign scores to interactions detected through Mass Spectrometry, Sowa and colleagues (*11*) unearthed an astonishing number of false positives in a list of 2,553 interactions, narrowing it down to 751 interactions. Examination of the list of interactions among human mitogen activated protein kinases (MAPKs) using the same procedure reduced the initial list of 2,000 interactions down to 641 interactions (*12*). The problem of false negatives (unidentified true binary protein interactions) requires the development of more precise procedures.

Detection of novel protein complexes through high-throughput studies is very dependent upon method's requirements and procedural characteristics. Key among the methodological limitations are: (i) the nature (Mode) of the interaction: whether the interactions are transient or permanent; (ii) the physiological conditions under which such interactions occur; (iii) the algorithms utilized for assigning scores during the identification of interaction complexes; (iv) the types of proteins identified--for example, interaction of plasma membrane and extracellular proteins versus those in the cytosol; and (v) Interactions between proteins may not be conserved in evolutionarily related organisms. Can these limitations be tackled?

In this review, we will discuss many of these aspects and propose future directions in the identification of protein–protein interaction complexes as well as the evolutionary dynamics these complexes undergo to enable the emergence of novel ecological adaptations. We first discuss alternative computational methods that present the potential to identify protein–protein interactions. Then, we elaborate on the applicability of these methods in two differentiated biological scales: the interactions between ligands and receptors and the interactions between biological systems.

MOLECULAR EVOLUTION CORRELATES WITH PROTEIN INTERACTIONS

Proteins are linked at the genetic level as well as at the functional level. Changes in one protein may therefore exert magnified effects in its interacting partner. These effects could disrupt cell viability when highly connected proteins (hubs) are involved. How strong of an evolutionary constraint is the link between two particular proteins?

Whole genome analyses have shown that more connected proteins are subjected to stronger selective constraints (they evolve slower) because they are more functionally constrained (*13, 14*). Moreover, interacting proteins tend to show similar levels of evolution (*13*). The similarity in the levels of evolution between interacting proteins is independent of the similarity in their respective number of interactions (*13*).

Studies measuring the correlation between the number of interactions for a protein and its evolutionary rate yield conflicting results in eukaryotes (*15, 16*) and in the bacterium *Escherichia coli* (*17*) as this correlation tends to vanish in response to the nature of the data set considered. Many confounding factors contribute to make the relationship between rates of evolution and interaction number vague: genes with higher expression levels, stronger codon bias, and shorter codon sequence tend to be highly constrained (*18–21*). Moreover, regulatory genes have been shown to evolve faster than structural genes (*22, 23*) and genes codifying for proteins acting downstream of particular pathways evolve slower than those acting upstream of these pathways (*24*). These confounding factors make it difficult to identify coevolution of protein-coding genes due to genuine interaction between their proteins.

If coadaptation were a general phenomenon between interacting proteins then detecting molecular coevolution could expose protein interactions. But exactly what is coevolution and how could we distinguish it from stochastic covariation? Ehrlich and Raven settled on the term coevolution from the ecological perspective when studying reciprocal evolutionary changes between butterflies and plants (*25*). Thompson (*26*) used this term in its widest meaning to refer to coevolution between species or populations. Following his definition, Thompson used the term "coevolution" to describe the correlated and consequential change of two populations, where changes in one population exert changes in the other in a reciprocal manner. This reciprocal change is itself under selection so that evolution of one population is highly dependent on the evolution of the other owing to the direct biological interaction between both populations.

In this review we consider coevolution as a term that generally refers to the evolutionary process in which a heritable change in the features of one entity establishes selective pressure for a change in another entity. These entities can span many different levels of complexity as long these levels are heritable and under selection: from nucleotides to amino acids to proteins to organisms and as high as ecosystems.

The concepts and methodologies developed to understand the coevolution of species and communities can be extrapolated to unearth coadaptation processes at the molecular level between genes and proteins. Proteins interacting through a set of amino acids in the interface is a clear example of coevolution, in which precise complementary structural conformations is crucial to maintain the interactions between the proteins. For instance, two proteins interacting in the cell through a specific set of amino acids are likely to present a coadaptation process: mutations in one of the proteins that interrupt conformation of one protein need to be compensated by a compensatory mutation at the interacting amino acid sites in the other protein to recover the structural complementarity between the two proteins, therefore generating a coevolutionary dynamic (Fig. 1). Although the concept is straightforward to conceive, coevolutionary dynamics could also be generated among amino acid sites that do not interact due to historical reasons or to stochastic processes (Fig. 1). Arguably, disentangling coevolutionary dynamics caused by interactions from those resulting from other dynamics will be instrumental in a method for *in silico* prediction of protein–protein interactions and protein complexes.

Current methods to identify coevolution between protein sequences suffer from serious limitations that hamper an accurate association of coevolution with functional/physical interactions. Is there a hope for the accurate computational prediction of protein complexes? Is the analysis of coevolution a mere complement to experimental approaches or does it have the potential of being an alternative? In this review we attempt to answer these questions exploring how promising computational biology is in defining organismal interactomes.

MOLECULAR COADAPTATION: MILLIONS OF YEARS OF TRIAL-ERROR EVOLUTION AS A MEANS TO IDENTIFY INTERACTIONS

The nature of interactions and the necessity for the coordinated change between interacting entities span several levels. In particular, protein functions are dependent upon their three-dimensional structure, which is the result of the complex atomic interactions between its components (amino acids) through energetically optimized processes (folding). Changes in these interactions would consequently lead to substantial changes in protein folding and ultimately affect protein structure and function. Owing to its stochastic occurrence, mutations are usually deleterious and only viable if compensated for by other mutations—a process termed coadaptation. Amino acids, therefore, ought to constrain one another (coevolve), as has been frequently demonstrated (for example, refs. *27–30*).

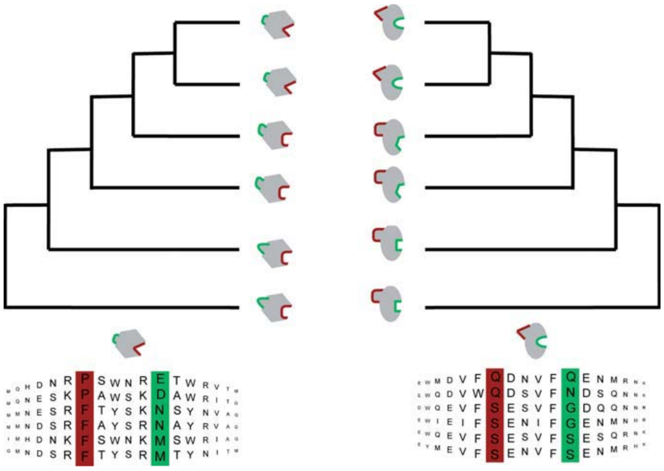
M Q H D N R P S W N R E T W R V T M
M Q N E S K P A W S K D A W R I T G
M M N E S R F T Y S K N S Y N V A G
M M N D S R F A Y S R N A Y R V A G
I M H D N K F S W N K M S W R I A G
Q M N D S R F T Y S R M T Y N I T M
E W M D V F Q D N V F Q E N M R N K
E W Q D V W Q D S V F N D S Q N N K
D W Q E V F S E S V F G D Q Q N N K
E W I E I F S E N I F G E N M N H K
E W Q E V F S E S V F S E S Q R N K
E Y M E V F S E N V F S E N M R H K

Figure 1. Coadaptation and historical correlation contribute to the coevolution of interacting proteins. We represent two interacting proteins using cartoons to facilitate visualization (oval and hexagon shapes, respectively). Each of the proteins has a phylogenetic history represented by their corresponding topologies, as shown. In each of the protein cartoons we represent two regions, one labeled in green and another in red. The green regions coevolve historically between the proteins—they present the same phylogenetic pattern and therefore the same evolutionary variation profile. These two sites, however, do not always present compatible geometries—as they do not form complementary surfaces. The red-labeled regions are involved in physical interaction between both proteins and are the ones coevolving functionally (they geometrically complement one another across the phylogeny). For the sake of simplicity we represent each of the coevolving regions with an amino acid site in a multiple sequence alignment for each protein; letters represent one-letter amino acid code and they increase in size as they approach the site involved in a coevolutionary relationship. Green columns correspond to the green region in the protein cartoons and red columns to the red region in the protein cartoons. It is noticeable that important changes in amino acid properties in the red column of one protein are always corresponded by important ones in the red column of the other protein, while this is not the case for the green columns.

Coevolution between amino acids, both within and between proteins, is a fundamental concept to understand evolution and evolvability because changes in one amino acid site of a protein may change the fitness landscape of its interacting amino acid partner. Coevolution is also fundamental at the interactome level in which interacting proteins ought to coevolve due to their reciprocal selective constraints (Fig. 1).

Coevolution has been extensively used to identify protein–protein interactions, with the underlying assumption being that interacting proteins coevolve. Methods to identify coevolution can be broadly classified in two categories: methods based on evolutionary distance information and tree-based methods. Methods based on evolutionary protein sequence distances compare the patterns of variation of amino acids between two proteins across a particular evolutionary scale. Homologous proteins from different organisms are matched so that equivalent amino acids within the sequence are aligned in a column. The variation along this column is then quantified by calculating the mutual information content (*27, 28, 31*). The variation can also be quantified by estimating the pairwise amino acid distance using probabilistic approaches and then, the correlation between the distance patterns between amino acid columns from the same or different proteins (*29, 32, 33*).

Methods based on the similarity of evolutionary trees work under the assumption that interacting proteins exert selection forces reciprocally upon one another and they should consequently present similar phylogenetic profiles (*34*). Recently, two methods have substantially improved the sensitivity of detecting protein–protein interactions through coevolutionary approaches. Juan et al. (*35*) extended these approaches by taking into account the coevolutionary force of a protein with the entire interactome of the bacterium *Escherichia coli*. Tillier and Charlebois (*36*) have analyzed the human interactome through a novel approach that does not assume conserved coevolutionary patterns throughout evolution. This method identifies the common distance submatrix in a pair of distance matrices quantifying the strength of coevolution through this matrix (*36*).

These methods have shown relative success in identifying protein–protein interactions using the molecular evolutionary dynamics but much remains to be done in this field. In particular, methods need to provide a cut-threshold coevolution value that could allow answer the question: are proteins "A" and "B" interacting? However, methods developed so far provide the first steps in paving the way for a novel concept of coevolution that extends beyond particular cases to embrace entire systems: including coevolution in metabolic pathways, interactomes and biological systems.

COEVOLUTION AND COADAPTATION OF MOLECULES: A CASE STUDY

The interaction between ligands and receptors is one example in which molecular variations in one protein is inextricably linked to variations in its interacting partner, in a way similar to the one depicted in Fig. 1. This results in an enormous conservation of important interactions between molecules through their evolutionary link (coevolution), particularly between molecules that mediate the interaction of the cell with its environment. Additionally high throughput methods to identify these interactions have not been developed to the same extent as those for intracellular ones. In this review we discuss two particular cases where coevolution may aid in understanding the ecological adaptation of organisms. In one case we comment on the necessity for the receptors in the extra-cellular region of cell membranes to coadapt with their ligands. In the second case we discuss on the ecological aspects of coadaptation of biological systems as a result of their molecular interactions.

EXTRACELLULAR PROTEIN INTERACTIONS ARE EVOLUTIONARILY CONSERVED

Among multicellular organisms cell surface and secreted (CSS) molecules play an essential role in mediating the communication with neighboring cells or the environment and trigger cellular responses to the environment. A vast number of processes inherent to a cellular communication are mediated by CSS proteins including basic responses such adhesion, attraction, repulsion etc. Not surprisingly, CCS interacting partners that mediate these processes are evolutionarily conserved. This conservation can be illustrated, for example, at the midline of bilaterians, an imaginary line that establishes the anteroposterior axis of symmetry and divides a bilaterian into a right and left side. A secreted protein from the leucin-rich repeat (LRR) family, Slit, is expressed at the midline establishing a repulsive barrier that prevents neuronal growth cones or cells from crossing the midline (*37*). These functions seem to be conserved from Platyhelminthes to vertebrates. The receptor for Slit, Roundabout (Robo), is an evolutionarily conserved type I transmembrane protein that belongs to the immunoglobulin superfamily. It is expressed in growth cones and upon binding Slit mediates its midline repulsive function preventing axons from crossing to the opposite side of the nervous system (*38*). Not only has this interaction been conserved during evolution but also the function of a presumably ancestral Slit/Robo interacting pair.

Once a successful interaction is established between CCS, the interacting partners can suffer duplication events and diversification resulting into large families. This phenomenon is easily noticeable among the Eph family of receptor tyrosine kinases (RTKs). The Eph receptors are found as early in evolution as poriferans. Its ancestral function might have been to direct cell segregation (*39*). They were subsequently adapted in eumetazoans to perform diverse functions in almost any cell-type in the organism. After a gene duplication some ephrins diversified into A-class and B-class ephrins, likely interacting with a single Eph receptor (*40*). Further duplication and diversifying events resulted in 14 Eph RTKs and 8 ephrins in mammals.

The examples of the Slit/Robo and Eph/ephrin complexes show how interactions that are functionally successful are conserved across evolution and how protein diversification can be influenced by protein interactions. The assumption supporting the idea that interacting CSS proteins should present stronger coevolutionary dynamics than noninteracting CSS proteins rests on the reasoning that interacting CSS proteins should be molecularly coupled as the entire system depends upon the integrity and conservation of these interactions. Slits and Robos, ephrins and Eph receptors, and in the same manner any other interacting pairs are consequently expected to present higher coevolutionary relationship. This strong coevolution between interacting proteins has been shown in a previous work where interacting proteins showed lower difference in their evolutionary rates than a set of randomly paired proteins (*13*).

DETERMINING INTERACTION SPECIFICITIES

The evolution of protein families certainly determines the interaction specificities within their members. Among chemokines and their receptors, the members of gene clusters usually show promiscuous binding due to a series of tandem gene-duplications that their ancestral genes suffered early in evolution (*41*). Likewise GPI-anchored and transmembrane (A and B class) ephrins show different specificity for EphA and EphB receptors respectively, that is, also dependent on their evolutionary history (*40*). Eph receptors contain an extracellular globular domain involved in ephrin interaction (*42*) that has a β-sandwich "jellyroll" folding topology (*43*). Ephrins interact through an induced-fit mechanism where they drive the formation of an extensive interaction surface within an unstructured channel present in the globular domain of the receptor. The specificity of the interaction resides in the spacing of certain bulky polar residues from the eprhin positioned against small polar residues inside the channel in the globular domain from the receptor that differ between A and B classes. This generates a key and lock arrangement where only ephrins and receptors from the same class can interact (*43*). It seems likely that, after an Eph receptor duplication, changes to the small polar residues present in the ephrin interacting surface of the receptor resulted in a better fit for one of the ephrins yielding both classes of receptors and ephrins. However, ephrins and Eph receptors present broad interaction promiscuity within each class. Currently, experimental methods to determine the residues responsible for specificity and promiscuity involve lengthy and, frequently, expensive techniques. Can we use coevolutionary analyses to identify specificity and its basis?

In theory, predicting the specificity of the interaction between two proteins can be possible using the values of correlation between amino acids belonging to the particular pair of interacting proteins as proxy. Since two interacting proteins form an interaction complex that is, essential for their function, both proteins should coevolve at their interacting surfaces (*44*). This coevolution can be manifested by the high conservation of the interface residues, where no evolutionary signal could be detected, or through the covariation at these residues presenting a coevolutionary signal (*44*). Using the appropriate computational and mathematical models, this signal could be used to determine the interacting pairs as well as the domains involved in the interaction since interacting domains between both proteins are likely to present a higher coevolutionary rate than non-interacting ones (*44*). In fact, this strong coevolution is mainly the result of a coadaptation relationship (deleterious-compensatory mutation dynamics) between interacting residues. This relationship has been shown to be particularly strong among clustered residues in proteins (*45*). Therefore, the fixing of mutations at one residue in the interface is subject to the fixing of compensatory mutations (conditional advantageous mutations) in the interacting residues, with both of these residues being entrenched in adaptive selection dynamics. This principle suggests that identification of residues that are critical for the interaction between proteins would be possible using computational-theoretical tools designed for identifying coadaptation processes.

However, an important limitation of most computational approaches to identify coadaptation processes is that they require sufficient evolutionary signal at the amino acid sites (amino acid variation) so that enough statistical power is provided for testing coevolution. To ensure a minimum evolutionary signal when comparing homologous genes a sufficiently sized set of orthologous genes (genes codifying for the same proteins in related species) should be used (*29*). In this respect, *Drosophila* would present an ideal test organism since 12 *Drosophila* genomes have been sequenced so far in addition to the genomes of the several species of mosquito, which are suitable as an outgroup. This number of sequences is well within the range of the number of sequences required to performing accurate coevolutionary analyses.

ORPHAN SURFACE MOLECULES AND UNSUSPECTED INTERACTIONS

Following an inverse rationale, one could exploit coevolution to determine novel protein–protein interactions that are vital for the cell and the organism function. Surface proteins play an essential role in an organism; as an example, within the nervous system, they control cellular migration, axonal guidance, branching, pruning, target recognition, synaptogenesis and synaptic plasticity just to name a few of the processes they regulate. There are abundant interactions among nervous system CSS described to date, however, there is a vast number of orphan CSS without known interacting partners. Different screens have highlighted the relevance of LRR family of proteins in the developing and adult nervous system (*46–48*). Among them, several proteins or families are still orphan such as the Elron or the Elfns (*46*). Likewise, there are undoubtedly, novel unsuspected interactions between non orphan CSS that have not been identified yet. Again, coevolutionary analysis at a whole genome level may help to identify or at least narrow down the search for possible interacting partners. How feasible is a genome-wide *in silico* identification of protein–protein interactions?

Computational analysis of coevolution, particularly when dealing with interprotein coevolution, is cumbersome because the number of combinations to be tested is substantial. For example, in a genome of *n* proteins, the number of possible coevolutionary analyses is [*n* x (*n*-1)/2]. In a small genome of 4,000 protein-coding genes the number of possible coevolutionary analyses is 7,998,000. To identify coevolution between two proteins, most methods measure the correlation in the amino acid variation patterns between every combination of two amino acids—one amino acid from one protein and another from the other protein. This means that for any two proteins of size *n* and *m* amino acids, respectively, we need to conduct $n \times m$ correlation analyses. For example, two proteins with a medium length of 300 amino acids would involve 90,000 calculus operations. In summary, a small genome of 4,000 genes would involve 7.2×10^{11} calculations of correlation, a computationally prohibitive task unless novel coevolutionary methods and models are developed specifically to target proteome-wide coevolutionary analyses.

Methods to identify coevolution have been hitherto devoted to the analysis of two specific proteins. Recently, however, an approach has been utilized to identify coevolution in the human proteome (*36*), using a simplified approach based on Mutual Information Content. Analysis of coevolution for the entire set of proteins in one biological entity or in two interacting biological species will require further models and methods feasible for proteome-wide analyses.

The *in silico* identification of protein–protein interactions would make it possible to answer fundamental questions in evolutionary biology and would advance forward in the understanding of the interaction between biological systems. These interactions are fundamental to the development of biological complexity and have been the basis for the emergence of eukaryotic cells. Arguably, development of appropriate methods to identify coevolution would pave the way to understanding biological complexity. One particular case of such complexity is the one established between prokaryotes and eukaryotes and that has taken place independently several times throughout evolution. What is the molecular basis for such a successful interaction? Could we use coevolutionary tools to identify the factors mediating the interaction between biological entities?

BIOLOGICAL INTERACTIONS AND THE MOLECULAR COEVOLUTIONARY PROCESS IN HOST-SYMBIONT SYSTEMS

Identifying interactions at the molecular level and the parameters governing their specificities is paramount to define the molecular basis for the biological interactions that justified the term "coevolution" in the first place—that is, the interaction between ecologically related species. Key among these interactions are the ones integrated into the broadly-termed species relationship "symbiosis," first defined by Anton De Bary in 1879 (*49*).

Within the context of the theoretical framework previously developed, we can find in the literature many potential candidate molecules mediating host-symbiont systems. Although we are unable to review herein all the biological processes involved in host-symbiont relationships, we can, however, provide some brush strokes on a few cases that are particularly prone to involve protein–protein coevolutionary dynamics between biological systems. For example, an important finding in host-symbiont interactions has been that the more we know the more difficult it is to distinguish between pathogenesis and mutualism because both kinds of bacteria share the same molecular mechanisms to establish the associations and evade host immune responses (*50*). Key among these molecular mechanisms are the secretion systems involved in the translocation of molecules to the extra-cellular space. Secretion systems play therefore an important role in the evolution of organism lifestyle, for example, by injecting effector proteins. Some bacterial endosymbionts utilize the type III secretion system (Inv/Spa genes) as a means to establish their symbiosis (*51*); other symbionts suppress the host immune defences or regulate virulence factors (*52*). Animal and plant bacterial pathogens rely both on types III (Hrp, Hop, Pop) and IV (Dot/Icm genes) secretion systems to infect their hosts (*53, 54*). Moreover, mutation of genes of the type III secretion system (Hrpg and Hrcv) can swap a plant pathogen into a nodule forming symbiont (*55*).

Mutualist bacteria of plants (legumes-Rhizobia mutualism) provide the host with nitrogen in exchange for photosynthetically fixed carbon. These bacteria express an array of signals, the Nod factors, that induce the development of the nodules in the plant roots by binding to host NFR proteins; the later binding further elicits the action of host receptor kinase proteins such as the SYMRK (*56*). The process of nodule formation, although morphologically different, involves the use of some molecules that are present as well in the ecologically widespread symbiosis between plants and arbuscular-mycorrhizal fungi (AMF). In addition, the plant-AMF symbiosis involves the exchange of nutrients (nitrogen and carbohydrates among others) with their plant hosts. During the symbiotic process of the plant and the fungi, the fungal symbiont expresses Myc factors, Nod factors, and Nitrogen metabolic genes; while the plant exudes a hormone to stimulate fungal metabolism and branching--the strigolactones. Moreover, this is followed by the expression of different plant species-specific product genes, for example, SYMRK, CASTOR, POLLUX, NUP85, NUP133, CCaMK, and CYCLOPS (*57*). Thus, rhizobiales bacteria and AMF share some molecular pathways aimed at the establishment of a morphological link with their hosts and the transfer of nitrogen; reciprocally, the process activates similar molecular pathways in the plant, although the final symbiotic structures are different. Recently, two new and independent symbiotic systems of plant and ants have been found to involve a third mutualist species of a fungus that provides nitrogen to the plants (*58, 59*). The study at the molecular level of these symbiotic systems might provide important insights in the evolution of symbioses that fix nitrogen in plants.

Other plant pathogen interactions, which are well documented due to their economical impact, are the plant-nematode interactions. Nematodes express genes involved in the degradation of the plant cell-walls such as β-1,4 endoglucanase (cellulose), pectate lyase, and polygalacturonase; other proteins, for example, chorismate mutase, act on the host cell formation (induction of syncytium and giant cells); while others disrupt host defences, for example, thioredoxin peroxidise, venom allergen-like protein, calreticulin (*60*). Some of the latter genes have a likely bacterial origin. The hosts express a battery of genes as well in response to the nematodes: host β-1,4 endoglucanase, polygalacturonase, pectin acetylesterase, AtSUC2 genes, PHAN, KNOX and members of the EREBP family transcription regulators, ENOD40, CCS52a (*60*).

The molecular interactions underlying host-symbiont relationship remain elusive. Nonetheless, the host immune system plays a decisive role in fuelling host-symbiont coevolution process by recognizing and resisting the invasion by the symbiont. In plants, there are two levels of immune response: first, the PRRs (transmembrane pattern recognition receptors) which are triggered by the presence of pathogen-associated components (MAMPS or PAMPs), such as chitin, flagellin, EF-Tu, β-glucan; and second, the intracellular expression of R genes (NB-LRR protein products), which recognize a broad range of pathogens (*61*). The molecular interaction between hosts and symbionts is a first step in gaining insight into biological complexity. Revealing the proteins involved in the interactions between systems and the molecular process whereby such interactions occur may shed light on the different outcomes of biological interactions—for example, mutualism versus pathogenesis. Not only will developing methods to identify molecular interactions allow understanding a phenomenon fundamental to the increasing biological complexity on Earth but will also make possible the identification of novel relationships among organisms yet unexplored. Thus, the refinement of bioinformatic tools to identify coevolutionary patterns will facilitate the design of more powerful experiments to analyse the relationships among species, and a means to either develop new strategies against pathogens or to handle economically important symbiotic events. This, however, remains a prospect for the foreseeable future.

ACKNOWLEDGEMENTS

This work was supported by the Science foundation Ireland, under the Research Frontiers Program (RFP: 10/RFP/GEN2685) and a grant from Ministerio de Ciencia e Innovación (BFU2009-12022) to M.A.F. J.P.L. is supported by the Principal Investigator Grant (07/IN.1/B913) and the Research Frontiers Program (08/RFP/NSC1617) from Science Foundation Ireland. The authors thank Prof. Herman Aberle, Dermot Harnett, and Ash Watson for their careful and critical reading of the manuscript.

REFERENCES

1. Albert, R., Jeong, H., and Barabasi, A. L. (2000) Error and attack tolerance of complex networks. *Nature* **406**, 378–382.
2. Goh, C. S., Milburn, D., and Gerstein, M. (2004) Conformational changes associated with protein-protein interactions. *Curr. Opin. Struct. Biol.* **14**, 104–109.
3. Edwards, A. M., Kus, B., Jansen, R., Greenbaum, D., and Greenblatt, J., et al. (2002) Bridging structural biology and genomics: assessing protein interaction data with known complexes. *Trends Genet.* **18**, 529–536.
4. Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M., et al. (1999) A generic protein purification method for protein complex characterization and proteome exploration. *Nat. Biotechnol.* **17**, 1030–1032.
5. Puig, O., Caspary, F., Rigaut, G., Rutz, B., Bouveret, E., et al. (2001) The tandem affinity purification (TAP) method: a general procedure of protein complex purification. *Methods* **24**, 218–229.
6. Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., et al. (2000) A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**, 623–627.
7. Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., et al. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA* **98**, 4569–4574.
8. Zhu, H., Bilgin, M., Bangham, R., Hall, D., Casamayor, A., et al. (2001) Global analysis of protein activities using proteome chips. *Science* **293**, 2101–2105.
9. Gavin, A. C., Aloy, P., Grandi, P., Krause, R., Boesche, M., et al. (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**, 631–636.
10. Krogan, N. J., Cagney, G., Yu, H., Zhong, G., Guo, X., et al. (2006) Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* **440**, 637–643.
11. Sowa, M. E., Bennett, E. J., Gygi, S. P., and Harper, J. W. (2009) Defining the human deubiquitinating enzyme interaction landscape. *Cell* **138**, 389–403.
12. Bandyopadhyay, D., Huan, J., Liu, J., Prins, J., Snoeyink, J., et al. (2010) Functional neighbors: inferring relationships between nonhomologous protein families using family-specific packing motifs. *IEEE Trans. Inf. Technol. Biomed.* **14**, 1137–1143.
13. Fraser, H. B., Hirsh, A. E., Steinmetz, L. M., Scharfe, C., and Feldman, M. W. (2002) Evolutionary rate in the protein interaction network. *Science* **296**, 750–752.
14. Hahn, M. W. and Kern, A. D. (2005) Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Mol. Biol. Evol.* **22**, 803–806.
15. Bloom, J. D. and Adami, C. (2003) Apparent dependence of protein evolutionary rate on number of interactions is linked to biases in protein-protein interactions data sets. *BMC Evol. Biol.* **3**, 21.
16. Pal, C., Papp, B., and Hurst, L. D. (2003) Genomic function: rate of evolution and gene dispensability. *Nature* **421**, 496–497; discussion 497–498.
17. Hahn, M. W., Conant, G. C., and Wagner, A. (2004) Molecular evolution in large genetic networks: does connectivity equal constraint? *J. Mol. Evol.* **58**, 203–211.
18. Duret, L. and Mouchiroud, D. (2000) Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. *Mol. Biol. Evol.* **17**, 68–74.
19. Pal, C., Papp, B., and Hurst, L. D. (2001) Does the recombination rate affect the efficiency of purifying selection? The yeast genome provides a partial answer. *Mol. Biol. Evol.* **18**, 2323–2326.
20. Rocha, E. P. and Danchin, A. (2004) An analysis of determinants of amino acids substitution rates in bacterial proteins. *Mol. Biol. Evol.* **21**, 108–116.
21. Ingvarsson, P. K. (2007) Gene expression and protein length influence codon usage and rates of sequence evolution in *Populus tremula*. *Mol. Biol. Evol.* **24**, 836–844.
22. Whitfield, L. S., Lovell-Badge, R., and Goodfellow, P. N. (1993) Rapid sequence evolution of the mammalian sex-determining gene SRY. *Nature* **364**, 713–715.
23. Rausher, M. D., Miller, R. E., and Tiffin, P. (1999) Patterns of evolutionary rate variation among genes of the anthocyanin biosynthetic pathway. *Mol. Biol. Evol.* **16**, 266–274.
24. Alvarez-Ponce, D., Aguade, M., and Rozas, J. (2009) Network-level molecular evolutionary analysis of the insulin/TOR signal transduction pathway across 12 *Drosophila* genomes. *Genome Res.* **19**, 234–242.

25. Ehrlich, P. R. and Raven, P. H. (1964) Butterflies and plants: a study in coevolution. *Evolution* **18**, 586–608.
26. Thompson, J. N. (1994) *The Coevolutionary Process*. University of Chicago Press, Chicago.
27. Atchley, W. R., Wollenberg, K. R., Fitch, W. M., Terhalle, W., and Dress, A. W. (2000) Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis. *Mol. Biol. Evol.* **17**, 164–178.
28. Gloor, G. B., Martin, L. C., Wahl, L. M., and Dunn, S. D. (2005) Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. *Biochemistry* **44**, 7156–7165.
29. Fares, M. A. (2006) Computational and statistical methods to explore the various dimensions of protein evolution. *Curr. Bioinform.* **1**, 207–217.
30. Lovell, S. C. and Robertson, D. L. (2010) An integrated view of molecular coevolution in protein-protein interactions. *Mol. Biol. Evol.*
31. Tillier, E. R. and Lui, T. W. (2003) Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments. *Bioinformatics* **19**, 750–755.
32. Pollock, D. D., Taylor, W. R., and Goldman, N. (1999) Coevolving protein residues: maximum likelihood identification and relationship to structure. *J. Mol. Biol.* **287**, 187–198.
33. Choi S. S., Li, W., and Lahn, B. T. (2005) Robust signals of coevolution of interacting residues in mammalian proteomes identified by phylogeny-aided structural analysis. *Nat. Genet.* **37**, 1367–1371.
34. Pazos, F. and Valencia, A. (2008) Protein co-evolution, co-adaptation and interactions. *EMBO J* **27**, 2648–2655.
35. Juan, D., Pazos, F., and Valencia, A. (2008) Co-evolution and co-adaptation in protein networks. *FEBS Lett.* **582**, 1225–1230.
36. Tillier, E. R. and Charlebois, R. L. (2009) The human protein coevolution network. *Genome Res.* **19**, 1861–1871.
37. Brose, K., Bland, K. S., Wang, K. H., Arnott, D., Henzel, W., et al. (1999) Slit proteins bind Robo receptors and have an evolutionarily conserved role in repulsive axon guidance. *Cell* **96**, 795–806.
38. Kidd, T., Bland, K. S., and Goodman, C. S. (1999) Slit is the midline repellent for the robo receptor in *Drosophila*. *Cell* **96**, 785–794.
39. Drescher, U. (2002) Eph family functions from an evolutionary perspective. *Curr. Opin. Genet. Dev.* **12**, 397–402.
40. Mellott, D. O. and Burke, R. D. (2008) The molecular phylogeny of eph receptors and ephrin ligands. *BMC Cell Biol.* **9**, 27.
41. Zlotnik, A., Yoshie, O., and Nomiyama H. (2006) The chemokine and chemokine receptor superfamilies and their molecular evolution. *Genome Biol.* **7**, 243.
42. Labrador, J. P., Brambilla, R., and Klein, R. (1997) The N-terminal globular domain of Eph receptors is sufficient for ligand binding and receptor signaling. *EMBO J* **16**, 3889–3897.
43. Himanen, J.-P. and Nikolov, D. B. (2003) Eph signaling: a structural view. *Trends Neurosci.* **26**, 46–51.
44. Pazos, F., Olmea, O., and Valencia, A. (1997) A graphical interface for correlated mutations and other protein structure prediction methods. *Comput. Appl. Biosci.* **13**, 319–321.
45. Davis, B. H., Poon, A. F., and Whitlock, M. C. (2009) Compensatory mutations are repeatable and clustered within proteins. *Proc. Biol. Sci.*
46. Dolan, J., Walshe, K., Alsbury, S., Hokamp, K., O'Keeffe, S., et al. (2007) The extracellular leucine-rich repeat superfamily; a comparative survey and analysis of evolutionary relationships and expression patterns. *BMC Genomics* **8**, 320.
47. Kurusu, M., Cording, A., Taniguchi, M., Menon, K., Suzuki, E., et al. (2008) A screen of cell-surface molecules identifies leucine-rich repeat proteins as key mediators of synaptic target selection. *Neuron* **59**, 972–985.
48. Hong, W., Zhu, H., Potter, C. J., Barsh, G., Kurusu, M., et al. (2009) Leucine-rich repeat transmembrane proteins instruct discrete dendrite targeting in an olfactory map. *Nat. Neurosci.* **12**, 1542–1550.
49. Bary, A. D. (1879) *Die Erscheinung der Symbiose*. Strassburg.
50. Dale, C. and Moran, N. A. (2006) Molecular interactions between bacterial symbionts and their hosts. *Cell* **126**, 453–465.
51. Dale, C., Young, S. A., Haydon, D. T., and Welburn, S. C. (2001) The insect endosymbiont *Sodalis glossinidius* utilizes a type III secretion system for cell invasion. *Proc Natl Acad Sci USA* **98**, 1883–1888.
52. Medina, M. and Sachs, J. L. (2010) Symbiont genomics, our new tangled bank. *Genomics* **95**, 129–137.
53. Sexton, J. A. and Vogel, J. P. (2002) Type IVB secretion by intracellular pathogens. *Traffic* **3**, 178–185.
54. Buttner, D. and He, S. Y. (2009) Type III protein secretion in plant pathogenic bacteria. *Plant. Physiol.* **150**, 1656–1664.
55. Marchetti, M., Capela, D., Glew, M., Cruveiller, S., Chane-Woon-Ming, B., et al. (2010) Experimental evolution of a plant pathogen into a legume symbiont. *PLoS Biol.* **8**, e1000280.
56. Radutoiu, S., Madsen, L. H., Madsen, E. B., Felle, H. H., Umehara, Y., et al. (2003) Plant recognition of symbiotic bacteria requires two LysM receptor-like kinases. *Nature* **425**, 585–592.
57. Parniske, M. (2008) Arbuscular mycorrhiza: the mother of plant root endosymbioses. *Nat. Rev. Microbiol.* **406**, 763–775.
58. Defossez, E., Djieto-Lordon, C., McKey, D., Selosse, M. A., and Blatrix, R. (2010) Plant-ants feed their host plant, but above all a fungal symbiont to recycle nitrogen. *Proc. Biol. Sci.*
59. Leroy, C., Séjalon-Delmas, J. A., Ruiz-González, M. X., Gryta, H., et al. (2011) Trophic mediation by a fungus in an ant-plant mutualism. *J. Ecol.* **99**, 583–590.
60. Williamson, V. M. and Gleason, C. A. (2003) Plant-nematode interactions. *Curr. Opin. Plant Biol.* **6**, 327–333.
61. Jones, J. D. and Dangl, J. L. (2006) The plant immune system. *Nature* **444**, 323–329.