Published online 12 November 2008 *Nucleic Acids Research, 2009, Vol. 37, Database issue* ***D629–D635***
doi:10.1093/nar/gkn853

EcID. A database for the inference of functional interactions in *E. coli*

Eduardo Andres Leon[1], Iakes Ezkurdia[1], Beatriz García[2], Alfonso Valencia[1,*] and David Juan[1]

[1]Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre–CNIO and [2]Computer Sciences Department, Universidad Carlos III de Madrid, Spain

Received August 22, 2008; Revised October 6, 2008; Accepted October 16, 2008

ABSTRACT

The EcID database (*Escherichia coli* Interaction Database) provides a framework for the integration of information on functional interactions extracted from the following sources: EcoCyc (metabolic pathways, protein complexes and regulatory information), KEGG (metabolic pathways), MINT and IntAct (protein interactions). It also includes information on protein complexes from the two *E. coli* high-throughput pull-down experiments and potential interactions extracted from the literature using the web services associated to the iHOP text-mining system. Additionally, EcID incorporates results of various prediction methods, including two protein interaction prediction methods based on genomic information (Phylogenetic Profiles and Gene Neighbourhoods) and three methods based on the analysis of co-evolution (Mirror Tree, *In Silico* 2 Hybrid and Context Mirror). EcID associates to each prediction a specifically developed confidence score. The two main features that make EcID different from other systems are the combination of co-evolution-based predictions with the experimental data, and the introduction of *E. coli*-specific information, such as gene regulation information from EcoCyc. The possibilities offered by the combination of the EcID database information are illustrated with a prediction of potential functions for a group of poorly characterized genes related to *yeaG*. EcID is available online at http://ecid.bioinfo.cnio.es.

INTRODUCTION

We present here a database, EcID (*Escherichia coli* Interaction Database), which includes specific features that facilitate the detection, and the assignment of reliability, to protein interactions and the generation of species-specific hypothesis on biological functions.

As specific features EcID integrates information from large-scale proteomic experiments (1,2), interactions directly extracted from text (3–7) and potential interactions obtained with various prediction methods (8–12) (Figure 1). Indeed the incorporation of prediction methods is one of the stronger features of EcID, a concept originally introduced for *E. coli* by the Indigo (13) database. More specifically EcID incorporates prediction methods based on co-evolutionary analysis that have been shown to provide consistent information complementary to other experimental and prediction methods [see, for example, (12)]. The prediction methods included in EcID are: *Phylogenetic Profiles* (PP) (8), based on the conservation of orthologous genes; *Gene Neighbourhood* (GN) (9), relating functions with the conservation of gene order; *Mirror Tree* (MT) (10), focused on detecting global evolutionary similarity among protein trees; *In Silico 2 Hybrid* (IH) (11) that detects pair of sets of orthologues with a high content of inter-protein co-evolving residues; and *Context Mirror* (CM) (12), a recently published method developed to detect sets of orthologues co-evolving in a highly specific way. The three of these methods, *MT*, *IH* and *CM*, based on co-evolutionary information are specific of EcID, and constitute one of the main values of the system that makes it different from any other server available, including other *E. coli*-specific databases, such as Bacteriome (14), an interesting initiative that includes data on PP and GN, but not of the co-evolution-related methods. The drawback is that these methods require laborious processing steps, including the calculation of multiple sequence alignments and the derivation of protein trees for every set of orthologues (Figure 2). This complexity has posed important challenges to the development and maintenance of the system.

The second characteristic that makes EcID different from other systems, such as STRING (15) or Prolinks (16) is the incorporation of *E. coli*-specific information.

*To whom correspondence should be addressed. Tel: +34 91 732 8000 (ext 2179); Fax: +34 91 732 8000; Email: valencia@cnio.es

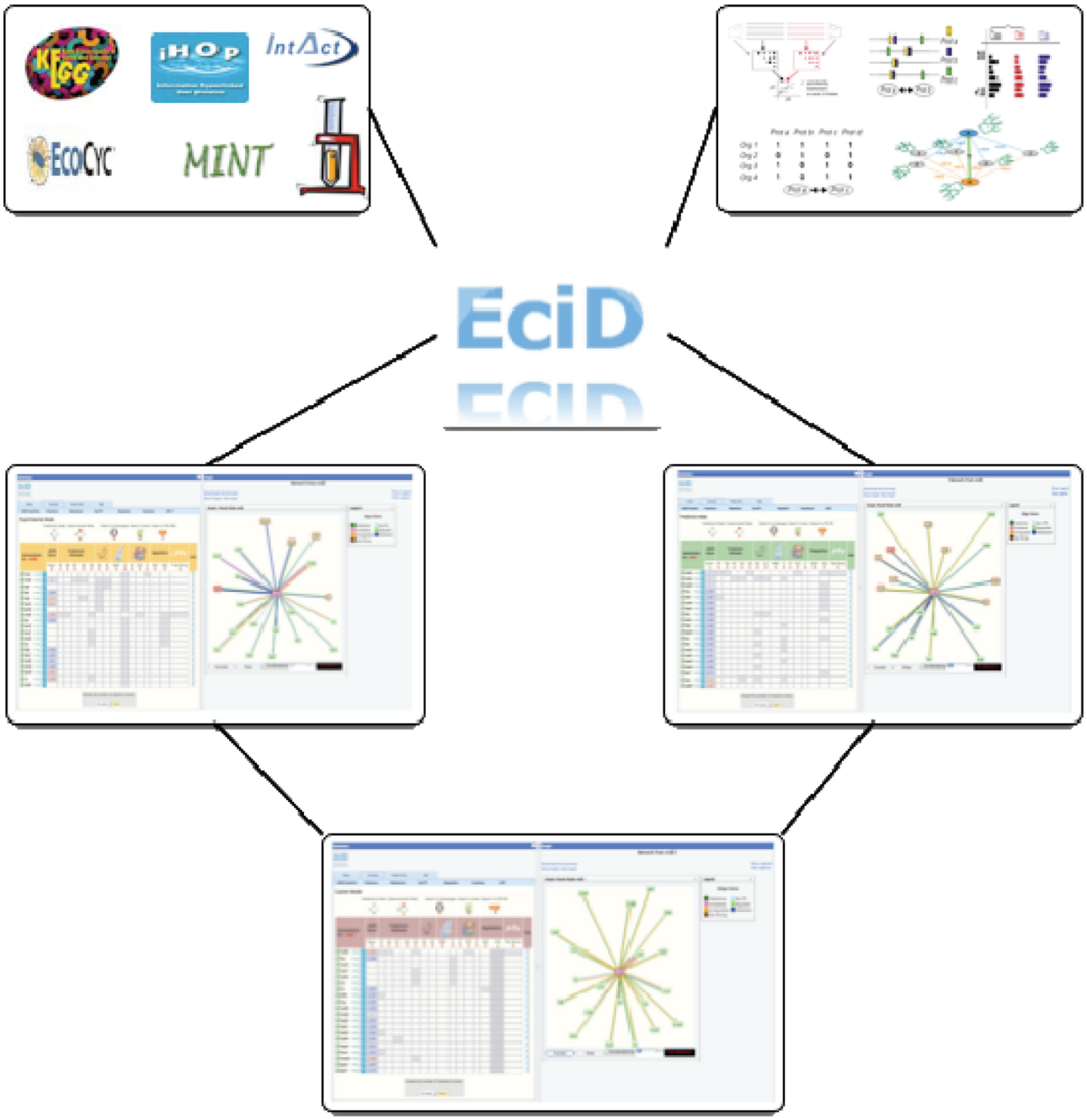

Figure 1. Overview of the data integration and display in EcID.

While these other databases are very useful multi-genome resources, with powerful general query system, they are not designed to provide information on specific organisms. For example, these databases do not include species-specific information, such as the manually curated complexes and pathways annotated in EcoCyc, a particularly useful repository of information about *E. coli*.

Indeed, we consider particularly interesting to develop an *E. coli*-specific system that will facilitate the work with collections of interactions in this model organism. This focus on *E. coli* and the integration of various predictions derived from methods based on co-evolution make of EcID a unique resource. In the case of *E. coli* there is clear unbalance between the vast amount of biochemical/biological information available, and the still relatively poorly studied interactome, where only two high-throughput experiments (1,2) have been published. It is precisely in this situation where EcID can be useful, since it is designed to complement the interaction studies with species-specific functional and evolutionary-derived information. In other words, the combination of information on interactions (high- and low-throughput), functional associations and predicted interactions, in the EcID framework constitute a powerful platform for the

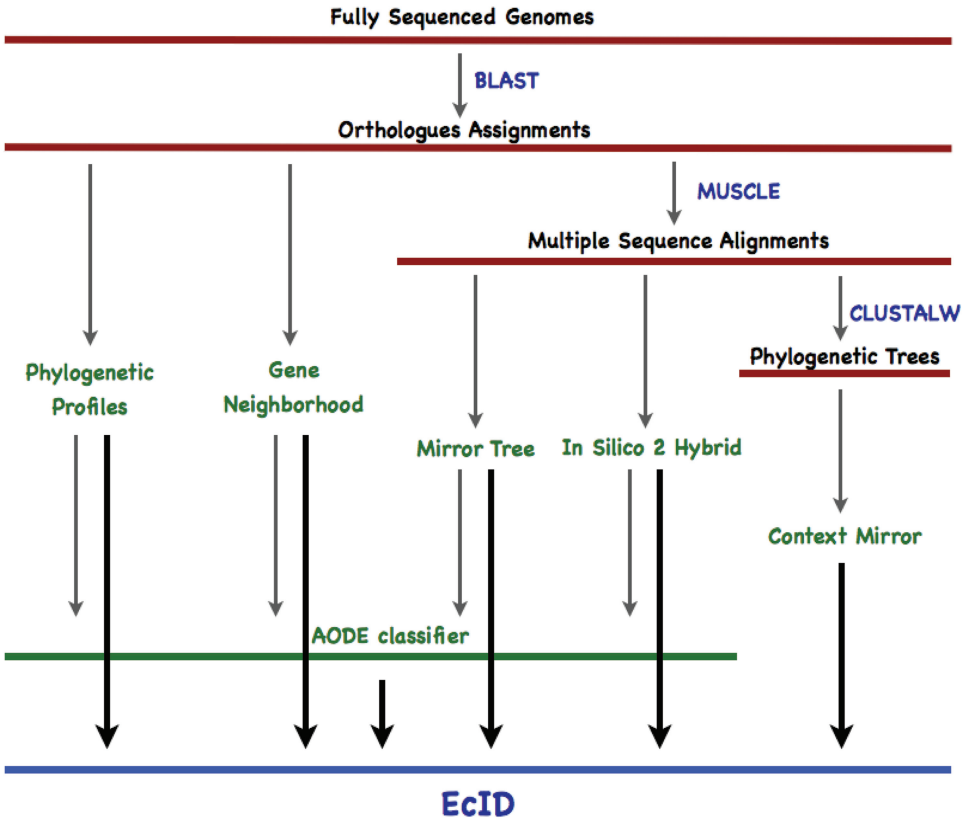

Figure 2. Outline of the data flow of interaction predictions in EcID.

generation of testable functional hypotheses. EcID includes two complementary navigation modes that prioritize the access to predicted, or experimental interactions (see Interface and database access Section). EcID also provides a 'Custom Mode' that allows focusing on interactions provided by specific data sources or computational methods. The various access modes allow combining the descriptive and exploratory power of the various data types provided by EcID.

DATABASE CONSTRUCTION

EcID is a database built on PostgreSQL, an open source relational database system particularly suited to provide fast access to large amounts of data. UniProt is used as reference from which some basic information is extracted including protein name, accession number UniProt ID and protein description.

The EcID is filled with information imported from external data sources (1–7), as well as inhouse calculations from different methods (8–12) (Figure 1). The complete task is complex given the diversity and heterogeneity of data. The acquisition of external data requires extensive data retrieval, parsing and re-formatting to a common internal format for every data type included in the database. The generation of the internal data from various prediction methods requires the retrieval of fully sequenced genomes from the EBI databases (17) (ftp://ftp.ebi.ac.uk/pub/databases/integr8/), extensive BLAST (18) searches, orthologues assignment using the 'Best BLAST Bidirectional Hits' (19) approach (we use blastp with an *e*-value threshold of 1E-5 calculated on a fixed database effective length of 10^8 and requiring 70% of both sequences to be aligned), and the building of multiple sequence alignments for every set of orthologues using MUSCLE (20), the calculation of neighbour joining trees using CLUSTALW (21) and the execution of each of the prediction methods (Figure 2).

EcID's web interface is based on PHP (Php 4.3), a dynamic language that facilitates database querying. The interface was developed using JavaScript and incorporates a Java applet for a graphical visualization of the protein network.

DATABASE CONTENTS AND COMPUTATIONAL METHODS

External databases imports and types of proteins functional interactions

The following classes of functional associations were imported from the external databases:

- Physical interactions derived from low-throughput experiments were extracted from the manual expert annotations of the IntAct (3) (http://www.ebi.ac.uk/intact/) and MINT (4) (http://mint.bio.uniroma2.it/mint/) databases.

- Proteins forming part of complexes, for which we established two sub-classes:
- High-quality complexes extracted from EcoCyc (5). This information is based on manual curation of the scientific literature and represents a set of well-known complexes.
- Protein complexes extracted from the publications by Butland *et al.* (1) and Arifuzzaman *et al.* (2). These high-throughput pull-down experiments provide extensive information of similar nature to the set of complexes extracted from EcoCyc, even if it is expected to be of lower quality.
- Regulator-regulated gene associations. We extracted the transcription regulatory data contained in EcoCyc (5) and established functional links between each transcriptional regulator and its corresponding regulated genes.
- Set of co-regulated genes. Based on the same type of information as the previous one, we established functional associations among those proteins that are coded by genes regulated by the same transcription factor.
- Proteins involved in the same biochemical pathway. These functional associations are extracted from the KEGG (6) and EcoCyc (5) databases. We considered all the proteins assigned to the same pathway to be functionally associated by pairs, even if there is not direct physical interaction between them.
- Interactions directly extracted from the literature using text-mining techniques. We include the protein name interactions obtained from iHOP (7) for *E. coli* proteins using the iHOP web service functionality (22). In this case, protein interactions are defined as those in which the corresponding names appear in the same sentence of a PubMed abstract.

Computational methods and evolutionary associations

EcID contains the results of five protein–protein interaction prediction methods. Essentially these methods are based on the detection of different types of evolutionary relationships in a set 227 prokaryotic fully sequenced genomes. The methods included in EcID are:

- IH method (11): this method predicts interacting partners by detecting pairs of co-evolving residues between different proteins. These residues are detected as pairs of positions showing correlated mutational behaviours in paired multiple sequence alignments of orthologous sequences. As in the original implementation (11), we calculate a score that represents the relative strength of inter-protein with respect to the potential intra-protein correlated mutations. The rank lists of possible interactors for the two proteins in the interaction pair are retrieved according to this value. The final score of the interactions is based on comparing the order of proteins in those two ranked lists of potential interactors (the technical details of the method are provided in the EcID web pages)
- MT method (10,23): MT is based on the observation of the similarity of phylogenetic trees of co-evolving protein families. In EcID, we include predictions obtained following the original article (10) where evolutionary similarities are calculated as Pearson's correlation coefficients between sequence distance matrices of paired multiple sequence alignments of orthologous proteins. Sorted lists of correlation coefficients for a given pair of proteins are used to calculate a new global score as in IH method.
- CM method (12): this recently published method proposes the prediction of protein interactions by analysing protein co-evolution at a proteome-wide level. This strategy is based on the use of the global network of protein evolutionary similarities to improve the detection of evolutionary parallelisms between proteins, followed by the analysis of the resulting co-evolutionary network in order to detect specific evolutionary similarities. CM method builds all the vectors (co-evolutionary profiles) containing the evolutionary similarities between a protein tree and the protein trees derived for every sequence of the reference organism, *E. coli* in this case. The co-evolutionary profiles are compared using Pearson's correlation coefficients. The specificity of co-evolution between two proteins is calculated by taking into account the potential influence of other proteins that might be influencing the co-evolution of those two proteins. The influence of these potential competing protein pairs is measured in terms of partial correlations between the corresponding co-evolutionary profiles. The scores shown in the system are the ones corresponding to partial correlation values where a maximum of 10 proteins are considered to be able to influence the co-evolution of each protein pair.
- GN method (9,19): this method is based on the organization of functionally related proteins in proximal regions of bacterial genomes, such as operons. This relation becomes more relevant when the neighbourhood relation is conserved in different species (24). One of the main limitations of this method is that it is only directly applicable to bacteria where the relation between transcription gene order and function is well established. In EcID, genes are considered functionally related if their orthologues in the other 227 genomes are closer than 300 bp.
- PP method (8,25): this method starts by building the patterns of presence or absence for every gene in a set of organisms, that is, determining if there is an orthologue of a given protein in every organism available. Similarity of these patterns or PP is interpreted as indicative of the need for both proteins to be present in order to perform a given function, even if it does not necessarily imply a physical relation between them. The main limitations of this approach are that it can only be applied to complete genomes and that it contains little information about essential proteins that tend to be present in most organisms. EcID defines the similarity between PP in terms of Euclidean distances between the presence/absence vectors.

A Bayesian-based global score

In order to provide a global score to combine the predictive power of the prediction methods described above,

we have integrated in EcID a Bayesian-based classifier (26). This method includes the scores of five different prediction methods (PP, GN, Gene Fusion (27,28), MT and IH), together with information on a number of simple sequence features. The classifier was trained to predict the type of protein functional associations contained in EcID (derived from regulation data, biochemical pathways, protein complexes and protein physical interactions; see 'External databases imports and types of proteins functional interactions' section) and it outperformed individual methods in the detection of functional association when compared in an independent test. More in detail the classification algorithm is based on the 'Averaged One-Dependence Estimators' (AODE) algorithm (29), a modification of the naïve Bayes (NB) method aimed to deal with inadequate independence assumption. This algorithm is computationally efficient and has obtained highly accurate predictions in other classification tasks. Furthermore, AODE provides quantitative probability estimates that can be used as a reliability measure associated to each predicted pair. This probability is used as a global confidence score and all the entries in EcID's 'Prediction Mode' are sorted according to this value (see the Interface and database access section).

INTERFACE AND DATA ACCESS

EcID allows free text queries on the UniProt entries stored in the database, including protein and gene names, UniProt IDs, accession numbers and protein functional description. Alternatively, the user can perform a BLAST (18) sequence search against the protein sequences stored in EcID, from which the desired entry can be selected.

Three navigation modes with the same overall interface are available in EcID:

- The 'Experimental Mode' gives access to experimentally validated associations ranked according to the reliability of the different data sources. EcID ranks experimental sources in the following order:

  Low-throughput experiments = Manually curated complexes > Metabolic pathways > Text mining > Regulation-related information > High-throughput experiments > Predicted interactions.
  In this mode, predictions can be considered as additional information about an experimentally detected interaction or as additional support for those associations derived from the less reliable experimental sources.

- The 'Prediction Mode' ranks the associations according to the relevance of the interactions calculated with the AODE's global score (29), and represents the amount of evolutionary-based information behind the predicted functional associations. This mode shows only entries with AODE's predictive scores and allows users to navigate the more exploratory information contained in EcID. The information based on predicted information can provide interesting new clues on interactions but it is not exempt of errors, and it has to be used with a well-informed user criterion.
- Finally, the system also allows users to select a specific type of data source or method to sort the potential interactions, in this 'Custom Mode' that selection preference is kept for the entire session.

The navigation is done using the same interface for the three modes and it is easy to change between them without the need of performing new protein queries. The initial navigation page contains a summary table displaying the proteins related to the query protein, the type of relation and links to the rest of the related information (see the Database contents and computational methods section). This summary table provides links to UniProt and to the information about interactions, and possible functions, available in Medline and automatically extracted using iHOP web services.

The protein associations organized in the table facilitate the navigation of the protein relationships network that can be additionally carried out using the EcID graph network viewer, where proteins are represented as nodes and associations as edges coloured by the various classes of associations. The viewer allows the selective display of protein links, the retrieval of information on the represented interactions, and to navigate the interaction network.

Finally, for a given query it is possible to download the complete list of interacting pairs and the information associated to each one of them in formats fully compatible with Cytoscape (30), Excel and PSI-MI (31). The full set of associations and scores contained in EcID is also openly accessible.

A simple usage example

To illustrate the possibilities offered by EcID, we use the example of the *yeaG* gene. This is a poorly characterized gene annotated with 'unknown function'. The 'Prediction Mode' of EcID recovers a list of 14 possible interactors with very different confidence values. The top ranked hit is *yeaH*, another uncharacterized gene, which is predicted as the most likely interactor with four out of the five prediction methods (all except IH), and scored with the highest confidence value by the AODE system. Interestingly, according with the EcoCyc annotation imported in EcID, both genes are regulated by the *glnG and glnL* transcription factors, also known as *ntrC* and *ntrB* that form part of the nitrogen assimilation two-component system. This information is coherent with a potential organization of *yeaG* and *yeaH* in an operon as also suggested by predictions rendered by the GN method. Additionally, the relation between *yeaG* and *glnG* is also supported by the predictions of IH method and by the significant AODE's combined score.

The second putative interactor with *yeaG*, also with a significant AODE's confidence value, is *ycgB* also described in the database as of unknown function. This prediction is supported by all the prediction methods except GN and reinforced by the fact that the text-mining system (iHOP) detects the following sentence in

a Medline abstract: *'Ten insertions mapped in nine open reading frames of unknown function (yciF, yehY, yhjY, yncC, yjgB, yahO, ygaU, ycgB, and yeaG) appear to be novel members of the RpoS regulon'* (32). Although this article refers to *Salmonella*, it adds an independent clue to complement the predicted functional associations and the possible implication of these proteins in stress response.

The following three proteins predicted in EcID as *yeaG* interactors (still over the 0.95 save confidence range of AODE) bring additional useful information. These proteins are: *glnP* (part of the glutamine transport system permease and predicted to be related with *yeaG* by the PP method), *glnL*, detected with the IH method and *glnG*, detected with by the global score, even if none of the scores of the individual methods was sufficiently high. These predictions reiterate the relation described above between *glnG* and *glnL* (*NtrB*/*C* nitrogen assimilation two components system) and *yeaG*.

The information provided by EcID points to the implication of *yeaG*, and *yeaH* in the nitrogen metabolism. More concretely, *yeaG* and *yeaH* could be related to response to nitrogen starvation (which seems to be regulated by *rpoS*) or to the missing link between NtrB/C nitrogen scavenging system and *rpoS* starvation response proposed by some authors (33). Additionally, EcID provides indirect evidences about a putative *yeaGH*/*ycgB* relationship. The high-confident predictions by all the different methods suggest a stronger association of what can be deduced from the common *rpoS* regulation. It has to be pointed out that this analysis of the potential functional relations is greatly facilitated by the EcID unified representation of predictions by different methods, information directly extracted from the literature and information on genomic organization (e.g. regulatory relationships retrieved from EcoCyc).

Of course, this example has to be taken just as a suggestion to be followed by the experimental verification of the proposed functions, and here it serves to demonstrate some of the capabilities offered by EcID.

CONCLUSIONS AND FUTURE DIRECTIONS

EcID is a database intended to provide an integrated framework for comparison and assessment of information on *E. coli* functional and physical interactions. For this purpose, EcID imports and classify information coming from the main external data sources on protein functional interactions, i.e. KEGG, EcoCyc, iHOP, MINT, IntAct and High-throughput experiments, and a set of five different protein interaction prediction methods based on evolutionary information. The main value of EcID, and what makes it different of other systems, is the incorporation of these prediction methods. In fact, EcID makes available three different co-evolution-based methods in order to improve the chances of obtaining new interesting predictions. Additionally, EcID incorporates a global Bayesian-based score that facilitates the prioritization of the predictions.

The database incorporates a sufficiently informative web interface and a simple and intuitive visualization system.

There are several aspects of EcID that will be improved in future versions. We are actively incorporating other contrasted methods, in particular, those based on homology-driven interaction inference and gene fusion events. Although these results are available from other resources, their inclusion in EcID would provide the opportunity of combining their results with the ones of the evolutionary-based methods. We are also working in introducing information on sequence/structure domains given their relevant role in protein function and interactions. The new version will incorporate an AJAX-based improved graphical interface that will allow simultaneous visualization of several levels of interactions, and we are considering the development of a cytoscape-specific plug-in. One final point we would like to tackle in the future is the inclusion of information directly provided by the user, such as prediction parameters and lists of potential interactions.

ACKNOWLEDGEMENTS

We want to thank the contribution of Robert Hoffmann, who developed iHOP, Chris Sander for kindly hosting the iHOP computational infrastructure and Jose Maria Fernandez-Gonzalez for the maintenance of the iHOP-based web services. We would also like to thank Florencio Pazos for the continuous discussion on the prediction methods. This work is largely based on the infrastructure created by the Spanish Nacional Bioinformatics Institute (www.inab.org) a platform of 'Genoma España'.

FUNDING

EC [LSHG-CT-2003-503265 (BioSapiens) and LSHG-CT-2004-512092 (EMBRACE)]. Funding for open access charge: European Union.

Conflict of interest statement. None declared.

REFERENCES

1. Butland,G., Peregrín-Alvarez,J.M., Li,J., Yang,W., Yang,X., Canadien,V., Starostine,A., Richards,D., Beattie,B., Krogan,N. *et al.* (2005) Interaction network containing conserved and essential protein complexes in Escherichia coli. *Nature*, **433**, 531–537.
2. Arifuzzaman,M., Maeda,M., Itoh,A., Nishikata,K., Takita,C., Saito,R., Ara,T., Nakahigashi,K., Huang,H., Hirai,A. *et al.* (2006) Large-scale identification of protein-protein interaction of Escherichia coli K-12. *Genome Res.*, **16**, 686–691.
3. Kerrien,S., Alam-Faruque,Y., Aranda,B., Bancarz,I., Bridge,A., Derow,C., Dimmer,E., Feuermann,M., Friedrichsen,A., Huntley,R. *et al.* (2007) IntAct—open source resource for molecular interaction data. *Nucleic Acids Res.*, **35**, D561–D565.
4. Chatr-aryamontri,A., Ceol,A., Palazzi,L.M., Nardelli,G., Schneider,M.V., Castagnoli,L. and Cesareni,G. (2007) MINT: the Molecular INTeraction database. *Nucleic Acids Res.*, **35**, D572–D574.
5. Keseler,I.M., Collado-Vides,J., Gama-Castro,S., Ingraham,J., Paley,S., Paulsen,I.T., Peralta-Gil,M. and Karp,P.D. (2005) EcoCyc: a comprehensive database resource for Escherichia coli. *Nucleic Acids Res.*, **33**, D334–D337.

6. Kanehisa,M., Araki,M., Goto,S., Hattori,M., Hirakawa,M., Itoh,M., Katayama,T., Kawashima,S., Okuda,S., Tokimatsu,T. *et al.* (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **36**, D480–D484.
7. Hoffmann,R. and Valencia,A. (2004) A gene network for navigating the literature. *Nat. Genet.*, **36**, 664.
8. Gaasterland,T. and Ragan,M.A. (1998) Microbial genescapes: phyletic and functional patterns of ORF distribution among prokaryotes. *Microb. Comp. Genomics*, **3**, 199–217.
9. Dandekar,T., Snel,B., Huynen,M. and Bork,P. (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.*, **23**, 324–328.
10. Pazos,F. and Valencia,A. (2001) Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Eng.*, **14**, 609–614.
11. Pazos,F. and Valencia,A. (2002) In silico two-hybrid system for the selection of physically interacting protein pairs. *Proteins*, **47**, 219–227.
12. Juan,D., Pazos,F. and Valencia,A. (2008) High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. *Proc. Natl Acad. Sci. USA*, **105**, 934–939.
13. Nitschké,P., Guerdoux-Jamet,P., Chiapello,H., Faroux,G., Hénaut,C., Hénaut,A. and Danchin,A. (1998) Indigo: a World-Wide-Web review of genomes and gene functions. *FEMS Microbiol. Rev.*, **22**, 207–227.
14. Su,C., Peregrin-Alvarez,J.M., Butland,G., Phanse,S., Fong,V., Emili,A. and Parkinson,J. (2008) Bacteriome.org–an integrated protein interaction database for E. coli. *Nucleic Acids Res.*, **36**, D632–D636.
15. von Mering,C., Jensen,L.J., Kuhn,M., Chaffron,S., Doerks,T., Krüger,B., Snel,B. and Bork,P. (2007) STRING 7—recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res.*, **35**, D358–D362.
16. Bowers,P., Pellegrini,M., Thompson,M., Fierro,J., Yeates,T. and Eisenberg,D. (2004) Prolinks: a database of protein functional linkages derived from coevolution. *Genome Biol.*, **5**, R35.
17. Kersey,P., Bower,L., Morris,L., Horne,A., Petryszak,R., Kanz,C., Kanapin,A., Das,U., Michoud,K., Phan,I. *et al.* (2005) Integr8 and genome reviews: integrated views of complete genomes and proteomes. *Nucleic Acids Res.*, **33**, D297–D302.
18. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
19. Overbeek,R., Fonstein,M., D'Souza,M., Pusch,G.D. and Maltsev,N. (1999) Use of contiguity on the chromosome to predict functional coupling. *In Silico Biol. (Gedrukt)*, **1**, 93–108.
20. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
21. Chenna,R., Sugawara,H., Koike,T., Lopez,R., Gibson,T.J., Higgins,D.G. and Thompson,J.D. (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.*, **31**, 3497–3500.
22. Fernández,J.M., Hoffmann,R. and Valencia,A. (2007) iHOP web services. *Nucleic Acids Res.*, **35**, W21–W26.
23. Goh,C., Bogan,A.A., Joachimiak,M., Walther,D. and Cohen,F.E. (2000) Co-evolution of proteins with their interaction partners. *J. Mol. Biol.*, **299**, 283–293.
24. Tamames,J., Casari,G., Ouzounis,C. and Valencia,A. (1997) Conserved clusters of functionally related genes in two bacterial genomes. *J. Mol. Evol.*, **44**, 66–73.
25. Pellegrini,M., Marcotte,E.M., Thompson,M.J., Eisenberg,D. and Yeates,T.O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA*, **96**, 4285–4288.
26. García,B., Juan,D., Ezkurdia,I., Andrés-León,E. and Valencia,A. (2008) Optimizing Protein Network Prediction with 'Averaged One-Dependence Estimators' (in press).
27. Marcotte,E.M., Pellegrini,M., Ng,H., Rice,D.W., Yeates,T.O. and Eisenberg,D. (1999) Detecting protein function and protein-protein interactions from genome sequences. *Science*, **285**, 751–753.
28. Enright,A.J., Iliopoulos,I., Kyrpides,N.C. and Ouzounis,C.A. (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**, 86–90.
29. Webb,G.I., Boughton,J.R. and Wang,Z. (2005) Not so naive Bayes: aggregating one-dependence estimators. *Mach. Learn.*, **58**, 5–24.
30. Shannon,P., Markiel,A., Ozier,O., Baliga,N.S., Wang,J.T., Ramage,D., Amin,N., Schwikowski,B. and Ideker,T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.
31. Hermjakob,H., Montecchi-Palazzi,L., Bader,G., Wojcik,J., Salwinski,L., Ceol,A., Moore,S., Orchard,S., Sarkans,U., von Mering,C. *et al.* (2004) The HUPO PSI's molecular interaction format—a community standard for the representation of protein interaction data. *Nat. Biotechnol.*, **22**, 177–183.
32. Ibanez-Ruiz,M., Robbe-Saule,V., Hermant,D., Labrude,S. and Norel,F. (2000) Identification of RpoS (sigma(S))-regulated genes in Salmonella enterica serovar typhimurium. *J. Bacteriol*, **182**, 5749–5756.
33. Peterson,C.N., Mandel,M.J. and Silhavy,T.J. (2005) *Escherichia coli* starvation diets: essential nutrients weigh in distinctly. *J. Bacteriol.*, **187**, 7549–7553.