Bioinformatics Advance Access published February 8, 2005

Protein function prediction using the Protein Link Explorer (PLEX)

Shailesh V. Date[1‡] *and Edward M. Marcotte*[1,2,*]

[1]*Center for Systems and Synthetic Biology,*
Institute for Cellular and Molecular Biology, and
[2]*Department of Chemistry and Biochemistry*
University of Texas at Austin,
1 University Station, A4800
Austin, Texas 78712-1064, USA

[‡]*Current Address: Center for Bioinformatics,*
1423 Blockley Hall, 423 Guardian Drive,
University of Pennsylvania, Philadelphia 19104, USA

[*]To whom all correspondence should be addressed

ABSTRACT

Summary: We introduce PLEX, a web-based environment that allows construction of a phylogenetic profile for any given amino acid sequence, and its comparison with profiles of ~350,000 predicted genes from 89 genomes, as a means of interactively identifying functionally-linked genes and predicting protein function. PLEX can be searched iteratively and also enables searches for chromosomal gene neighbors and Rosetta Stone linkages. PLEX search results are accompanied by quantitative estimates of linkage confidence, enabling users to take advantage of coinheritance, operon, and gene fusion-based methods for inferring gene function and reconstructing cellular systems and pathways.

Availability: http://bioinformatics.icmb.utexas.edu/plex

Contact: marcotte@icmb.utexas.edu

INTRODUCTION

Functional annotation of completely sequenced genomes has proved to be a formidable task, and large fractions of genes are as yet uncharacterized. Even in well-studied genomes, such as that of *Escherichia coli*, ~30% of the genes are annotated as being of unknown function. In the malarial parasite *P. falciparum,* ~60% of genes lack functional assignments (Gardner *et al.,* 2002). To better understand the biology of these organisms, associating functions, even general functions, with the uncharacterized genes is of paramount importance.

We have created the web-based **P**rotein **L**ink **EX**plorer (PLEX) system to allow interactive exploration of comparative genomics tools for inferring linkages between genes, specifically through the use of phylogenetic profile analysis (Pellegrini *et al.,* 1999), Rosetta Stone links (Marcotte *et al.,* 1999; Enright *et al*., 1999), and the inference of operons based upon the distance between adjacent genes (Salgado *et al.,* 2000). While pre-calculated linkages from phylogenetic profiles are available in the STRING (von Mering *et al.*, 2003) and Predictome (Mellor *et al.*, 2002) databases, tools for construction and search of user-input phylogenetic profiles are not widely accessible.

Using PLEX, a phylogenetic profile can be constructed from any given amino acid sequence, or even specified manually to reflect desired phylogenetic distributions, then compared with pre-calculated profiles of 350,111 proteins from 89 bacterial, archaeal, and eukaryotic genomes in the PLEX database. Information about Rosetta stone protein links and chromosomal gene neighbors is provided, and iterative searches are feasible. We anticipate the system will be of use to any biologist hoping to gain insight into a particular cellular system or to suggest genes responsible for an observed phenotype. For first time users of the system, a short guided tour is available at *http://bioinformatics.icmb.utexas.edu/plex/tour/*

IMPLEMENTATION

The PLEX system is based upon a mySQL relational database, storing gene sequences, chromosomal positions, pre-computed phylogenetic profiles and Rosetta stone linkages, accessible via PERL scripts from a web-based interface. PLEX first compares a user-entered amino acid sequence, using BLASTP under default settings, to ~350,000 predicted genes from 89 genomes to construct a phylogenetic profile, which is then compared versus those in the database to identify genes with a similar phylogenetic distribution from a user-specified genome. Similarity is evaluated by identifying the phylogenetic profiles with maximal mutual information to the query profile (calculated as in Date and Marcotte, 2003). Genes recovered in the search can be chosen as queries for new searches, in this manner exploring the space of co-inherited genes. Additional functional linkages are provided in the form of Rosetta Stone linkages that are identified and ranked via a statistical measure of confidence based upon the hypergeometric distribution (Verjovsky Marcotte and Marcotte, 2002). Gene neighbors are included when neighboring genes on the chromosome are closer than a specified number of nucleotides (typically 40 nucleotides, suggested by log-likelihood analysis of operon boundaries (Salgado *et al.,* 2000)).

Figure 1 illustrates the use of PLEX with two examples from the *Mycobacterium tuberculosis* genome – the reconstruction of all subunits of the urease enzyme, and reconstruction of the isoprenoid biosynthesis pathway, including several potential new

components, linked by PLEX to the main pathway. Starting from the amino acid sequence of the urease subunit UreA (red circle, Fig. 1A), a comparison with profiles of all genes in the *M. tuberculosis* genome revealed functional links with subunits UreB, UreC and UreG of the urease enzyme. Analysis of gene neighbors revealed the linkage between the UreB and UreA genes. Rosetta Stone linkages were found between the A, B & C subunits, the composite genes (AB, AC, BC) occurring in 8 different organisms. UreD and UreF were identified as operon partners by selecting UreB, UreC and UreG as queries; the distinct phylogenetic distributions of UreD and UreF are immediately evident in the figure (drawn at bottom). A similar analysis led to the reconstruction of the *M. tuberculosis* isoprenoid biosynthesis pathway (Fig. 1B), starting from the sequence of the IspG protein. Along with the known components (labeled by the steps they catalyze), four additional proteins are implicated in the isoprenoid biosynthesis pathway: Rv3583 (a putative transcription factor), Rv2869c (a zinc metalloprotease), Rv1012, and IdsB (a putative polyprenyl synthetase).

Specifying constraints on the phylogenetic profiles provides another route to finding genes associated with a pathway (e.g., as in Huynen *et al.*, 1998). For example, from a profile corresponding appropriately with flagellated and non-flagellated bacteria, PLEX returns genes involved in flagellar structure and biosynthesis. These genes can be chosen as queries for the determination of operon and Rosetta Stone linkages. In this manner, phenotypes can be associated directly with candidate genes and systems.

ACKNOWLEDGEMENTS

The authors acknowledge funding from the Welch Foundation (F-1515), Packard Foundation, NSF, NIH (GM067779-01), and Beckman/MURI.

REFERENCES

Date,S.V. and Marcotte,E.M. (2003) Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. *Nature Biotechnol.,* **21**, 1055-1062.

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.

Gardner *et al.* (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature*, **419**, 498-511.

Huynen,M., Dandekar,T. and Bork,P. (1998) Differential genome analysis applied to the species-specific features of Helicobacter pylori. *FEBS Letters* **426**, 1-5.

Marcotte,E.M., Pellegrini,M., Ng,H.-L., Rice,D.W., Yeates,T.O. and Eisenberg, D. (1999b) Detecting protein function and protein-protein interactions from genome sequences. *Science*, **285**, 751-753

Mellor, J.C, Yanai, I., Clodfelter, K.H., Mintseris, J. and DeLisi. C. (2002) Predictome: a database of putative functional links between proteins. *Nucleic Acids Res.*, **30**, 306–309.

Pellegrini,M., Marcotte,E.M., Thompson,M.J., Eisenberg,D. and Yeates,T.O. (1999) Assigning protein functions by comparative analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci.* USA, **96**, 4285-4288.

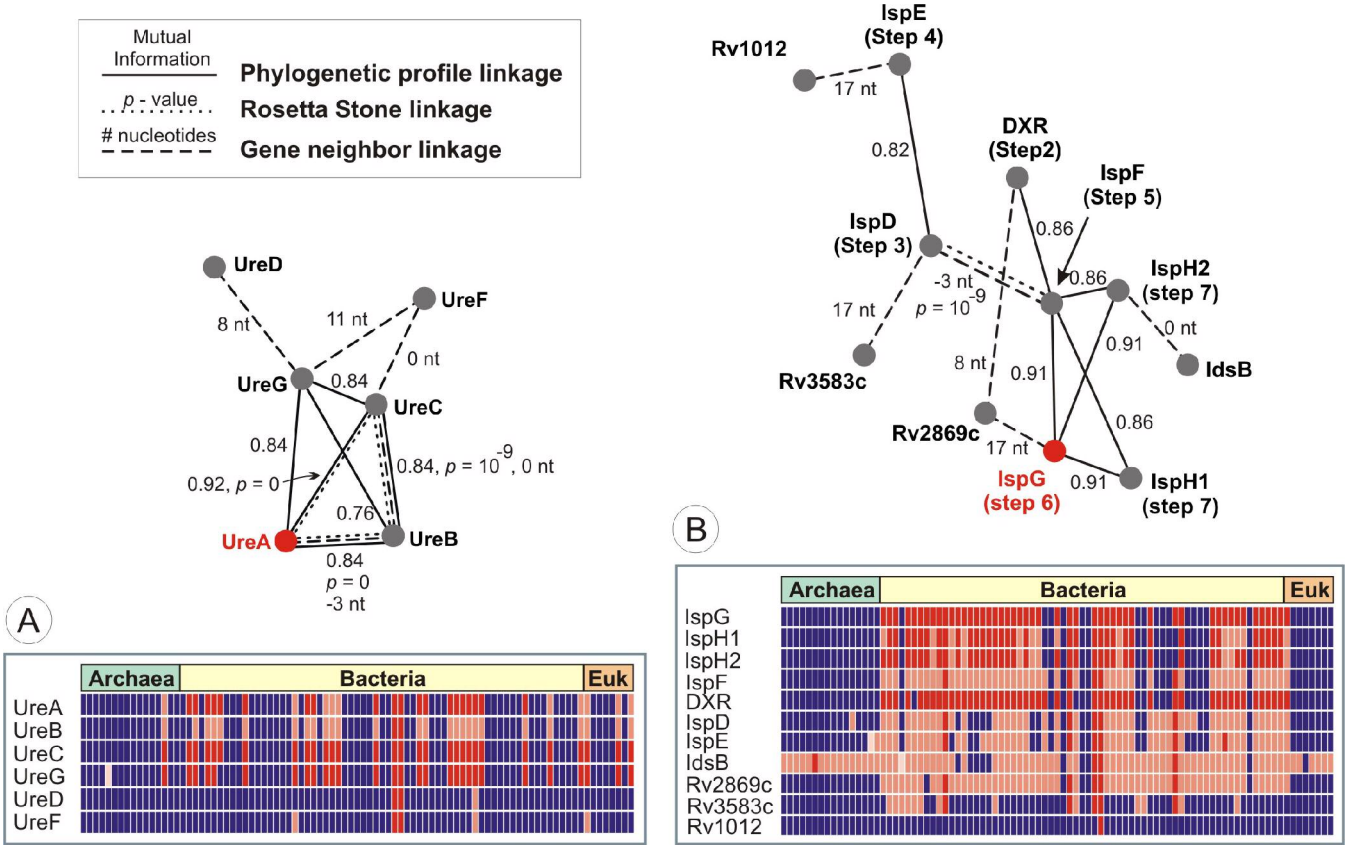
Salgado,H., Moreno-Hagelsieb,G., Smith,T.F. and Collado-Vides,J. (2000) Operons in Escherichia coli: genomic analyses and predictions. *Proc. Natl. Acad. Sci. USA,* **97**, 6652-6657.

von Mering,C., Huynen,M., Jaeggi,D., Schmidt,S., Bork,P. and Snel,B. (2003) STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res*. **31**, 258-261.

Verjovsky Marcotte,C.J. and Marcotte,E.M. (2002). Predicting functional linkages from gene fusions with confidence. *Applied Bioinformatics* **1**, 1-8

FIGURE LEGEND

Figure 1. Two examples of *Mycobacterium tuberculosis* protein systems (**A** - urease enzyme complex; **B** - isoprenoid biosynthesis pathway) reconstructed using the functional genomics methods available within PLEX (only significant links are drawn).


Mutual Information
Phylogenetic profile linkage
p - value
Rosetta Stone linkage
nucleotides
Gene neighbor linkage
UreD
UreF
8 nt
11 nt
0 nt
UreG
0.84
UreC
0.84
0.92, p = 0
0.84, p = 10^-9, 0 nt
0.76
UreA
UreB
0.84
p = 0
-3 nt
A
Archaea
Bacteria
Euk
UreA
UreB
UreC
UreG
UreD
UreF
IspE (Step 4)
Rv1012
17 nt
0.82
DXR (Step2)
IspF (Step 5)
IspD (Step 3)
0.86
-3 nt
p = 10^-9
0.86
IspH2 (step 7)
17 nt
0.91
0 nt
Rv3583c
8 nt
0.91
IdsB
0.86
Rv2869c
17 nt
IspG (step 6)
0.91
IspH1 (step 7)
B
Archaea
Bacteria
Euk
IspG
IspH1
IspH2
IspF
DXR
IspD
IspE
IdsB
Rv2869c
Rv3583c
Rv1012