URLs

BIOINFORMATICS

From genotype to phenotype: a shortcut through the library

A list of genes tells you little about their biological roles: understanding this generally requires time-consuming functional or comparative studies. A new method provides a shortcut on this path from genotype to phenotype — it uses a combination of comparative genomics and literature mining to predict the functions of large sets of sequenced genes.

Peer Bork and colleagues reasoned that if a group of species has a shared phenotypic trait, orthologous genes shared among the species are likely to be involved in the underlying biological process. Such genotype–phenotype correlations have been made in the past, but required initial manual collection of phenotypic information for each species, which is labour-intensive and might lead to biases in the phenotypes examined.

In the new study, this annotation stage was avoided by directly linking species with phenotypic information already available in the published literature. The authors linked 92 completely sequenced prokaryotic species with 172,967 nouns in MEDLINE abstracts (from the database compiled by the US National Library of Medicine). The nouns were grouped according to the species they matched up with, assuming that words that relate more frequently to a particular set of species are likely to be specific to a shared trait. For the same species, 11,026 orthologous gene sets were identified from the STRING (search tool for the retrieval of interacting genes/proteins) database, defining shared sets of genes.

The final stage was to look for correlations between the groupings of MEDLINE nouns and the groupings of shared genes. Bork and colleagues identified 2,700 significant associations between orthologous groups and trait words, which allowed them to relate 28,888 genes to at least one trait. Among these associations, many of the gene–phenotype associations were already known, confirming the validity of the method. However, many new discoveries were also made. For example, previously unknown associations were made between a group of metabolic genes and trait words linked to food poisoning.

Many of the other associations made in this study also linked genes to disease-related phenotypes, which could provide a valuable source of new drug targets. This skew towards clinically relevant phenotypes reflects the fact that pathogens are highly represented among fully sequenced prokaryotes, but as more sequences and more MEDLINE entries become available, this method should provide a way to link genes to a wide range of biological processes.

Louisa Flintoft

References and links

ORIGINAL RESEARCH PAPER
Korbel, J. O. *et al.* Systematic association of genes to phenotypes by genome and literature mining. *PLoS Biol.* **3**, e134 (2005)

WEB SITES
Peer Bork's laboratory: http://www-db.embl.de/jss/EmblGroupsHD/g_27.html
The STRING database: http://string.embl.de