PublisherInfo		
PublisherName	:	BioMed Central
PublisherLocation	:	London
PublisherImprintName	:	BioMed Central

Predicting protein interactions from sequence

ArticleInfo		
ArticleID	:	3543
ArticleDOI	:	10.1186/gb-2000-1-1-reports009
ArticleCitationID	:	reports009
ArticleSequenceNumber	:	34
ArticleCategory	:	Paper report
ArticleFirstPage	:	1
ArticleLastPage	:	4
ArticleHistory	:	RegistrationDate : 1999–11–29 Received : 1999–11–29 OnlineDate : 2000–3–17
ArticleCopyright	:	BioMed Central Ltd2000
ArticleGrants	:	
ArticleContext	:	130591111

Paul Kellam

Abstract

Bioinformatic analysis has been used to identify gene-fusion events in complete genomes and thus infer protein interactions on the basis of sequence analysis alone.

Significance and context

As part of post-genomic research, extensive efforts are underway to analyze protein-protein interactions. The overall aim of such analysis is to identify proteins that are functionally related, for example proteins that participate in a common structural complex, a metabolic pathway or a biological process. Currently, this involves using a range of biochemical and molecular biology methods such as two-hybrid systems and co-immunoprecipitation. The labor-intensive nature of these studies and associated technical difficulties make these approaches daunting when considered on a whole-genome scale. This paper is one of the first to use purely computational methods of sequence comparison to predict protein-protein interactions, and represents a new use of bioinformatics in post-genomic research. It is known that certain protein families consist of joined (fused) protein domains resulting in a 'composite' (as defined by Enright *et al.*) full-length protein. The individual proteins that fuse to create such a composite protein are referred to as 'component' proteins. The authors hypothesize that, in the context of a complete genome, if a composite protein is uniquely similar to two or more component proteins in the complete genome of another species, then a fusion event has occurred and therefore, the component proteins are most likely to interact (either physically or functionally).

Key results

Using their algorithm the authors identified 215 component proteins in four query genomes, *Escherichia coli, Haemophilus influenzae, Methanococcus jannaschii* and *Saccharomyces cerevisiae.* The proteins were involved in 88 fusion events, of which 64 were unique. Ninety-four composite proteins were identified in four reference genomes, of which 17 were paralogous composite proteins, leaving 77 fusion cases. Only three multiple fusion events were observed. The validity of the technique was shown by the fact that 26 of the 64 component proteins identified as fused are known to be involved in the same protein complex or biochemical process.

Methodological innovations

The authors use a novel analysis method that compares each of the proteins encoded by a complete genome (the 'query genome') with each other using BLASTP to produce a matrix T. The query genome proteins are then compared with a reference genome, also using BLASTP, to produce matrix Y. A fusion detection algorithm is used to find when two or more individual proteins in the query genome that have no similarity to each other (by checking matrix T) exhibit similarity to a single protein in the reference genome (by checking matrix Y). Further details can be found in the Supplementary information to *Nature* **402**:85-90 to the paper.

Conclusions

The authors conclude that with the addition of new complete genomes and sufficient computational power this method will be a valuable addition to the tools for discovering functional relations among proteins.

Reporter's comments

This paper represents the first use of sequence comparison methods to identify protein interactions. Previous approaches have involved the study of subunit interfaces from a structural viewpoint to identify such interactions. The new method on its own is a valuable contribution to the needs of post-genomic analysis: it is estimated to have few false positives, and when combined with other methods of discerning functional interactions (such as correlated mRNA expression patterns) generates a considerable body of *in silico* data. This can be used to guide molecular biology to investigate the most likely protein interactions. A related paper using such a combined approach and a 'News and Views' feature covering both papers can be found in the same issue of *Nature*.

Table of links

Nature

Supplementary information to *Nature* **402**:85-90

References

1. Enright AJ, Lliopoulous I, Kyrpides NC, Ouzounis CA: Protein interaction maps for complete genomes based on gene fusion events. Nature. 1999, 402: 86-90. 0028-0836