SPRINT: Ultrafast protein-protein interaction prediction of the entire human interactome

Yiwei Li & Lucian Ilie*
Department of Computer Science
University of Western Ontario, CANADA

Abstract

- Protein-protein interaction (PPI) prediction – a fundamental problem in system biology
- Experimental methods (Y2H, TAP) are inaccurate and time and labor intensive
- Many computational approaches are proposed: sequence-based ones are very promising
- Current sequence-based programs are too slow
- SPRINT - a new sequence-based algorithm for PPI prediction
- More accurate than the leading sequence-based programs
- Orders of magnitude faster
- The only program that can effectively predict the entire human interactome

The SPRINT algorithm

Step 0: The Idea

Proteins similar with interacting proteins are likely to interact as well. SPRINT uses a complex algorithm to quickly evaluate the contribution of similar subsequences to the likelihood of interaction. The basic idea is illustrated in Figure 4 where blocks of the same colour indicate similar subsequences and (P1, Q1) is a known interaction. Each pair of blocks in (P1) and (Q1) (dashed line) increases the likelihood of interaction between proteins containing similar subsequences.

Step 1: Finding similar subsequences

Highly sensitive multiple spaced seeds are used for very fast and reliable comparison of similar subsequences (1 = match, = don’t care; e.g., 11****11***1). In addition to exact hits (Figure 5a), approximate hits (Figure 5b) are essential in achieving high sensitivity. The usual d-mers are replaced by $d$-mers (Figure 6). Retrieval operations are heavily used for speed.

Step 2: Post-processing similarities

Similar subsequences that appear too often are removed. In Figure 7 similarities are marked by lines (a) and positions with larger counts (5 or larger in the example) are removed (b).

Step 3: Scoring PPIs

The score for each protein pair is computed by adding the PAM120 (default) score of each similarity pair (dashed line in Figure 4) and then normalized by the product of the protein lengths.

Conclusions

- SPRINT is a more accurate and much faster sequence-based PPI prediction algorithm and tool
- Our goal is to make predicting the entire human interactome a routine task

Availability

The source code of SPRINT is freely available from http://www.csd.uwo.ca/faculty/ilie/SPRINT/ and the datasets and predicted PPIs from http://www.csd.uwo.ca/faculty/ilie/SPRINT/.

References


Acknowledgements

This work has been partially supported by a Discovery Grant and a Research Tools and Instruments Grant from the Natural Sciences and Engineering Research Council of Canada (NSERC) to L.I.