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

Yiwei Li & Lucian Ilie[*]

Department of Computer Science
University of Western Ontario, CANADA

[*]Corresponding author: `ilie@uwo.ca`

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 sequenced-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

Results

Dataset	All PPIs	Training	Testing	Website
Park and Marcotte	24,718	14,186	1,250	`www.marcottelab.org/` `differentialGeneralization`
Biogrid	215,029	100,000	10,000	`thebiogrid.org`
HPRD Release 9	34,044	10,000	1,000	`www.hprd.org`
InnateDB experimentally validated	165,655	65,000	6,500	`www.innatedb.com`
InnateDB manually curated	9,913	3,600	360	`www.innatedb.com`
IntAct	111,744	52,500	5,250	`www.ebi.ac.uk/intact`
MINT	16,914	7,000	700	`mint.bio.uniroma2.it`

Table 1: The datasets used for comparison. Each dataset is used to create three types of testing data, depending on whether both test proteins appear in training (C1), only one appears (C2), or none (C3).

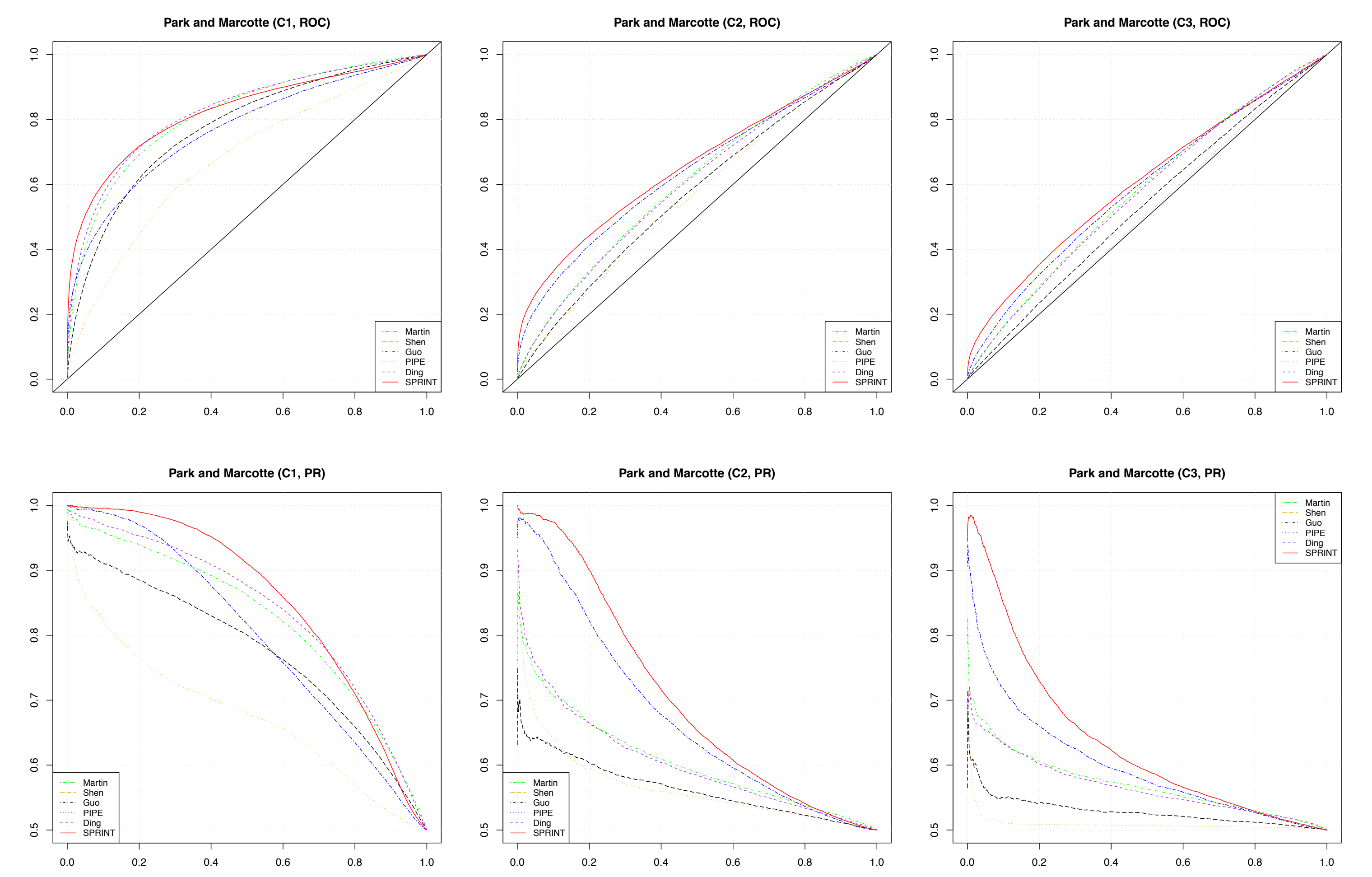

Figure 1: Comparison against leading programs: Martin's,[5] PIPE2,[6] Shen,[7] Guo,[3] and Ding[2] on Park and Marcotte[1] dataset: ROC (top) and PR (bottom) curves. All programs except PIPE2 and SPRINT use machine learning. Note the increasing difficulty from C1 to C3.

Table 2: Comparison on all seven datasets for high specificity, which matters most for PPI prediction. Each C1-3 average is computed from all seven corresponding C1-3 datasets.

Dataset	Specificity	Sensitivity				Precision				F1-score			
		Martin	PIPE2	Ding	SPRINT	Martin	PIPE2	Ding	SPRINT	Martin	PIPE2	Ding	SPRINT
C1 average	99.95%	6.07	7.60	11.93	13.35	98.52	98.82	88.05	99.37	11.06	13.55	20.39	22.93
	99.90%	6.53	9.20	14.24	15.91	97.36	98.61	90.65	99.29	11.88	16.45	24.10	27.03
	99.50%	17.27	21.41	29.90	29.50	96.66	97.52	98.20	98.30	28.62	34.73	45.19	45.22
	99.00%	25.48	28.73	38.72	40.14	95.55	96.40	97.28	97.52	39.14	43.69	54.69	56.58
	95.00%	55.35	48.07	65.68	62.02	91.44	90.19	92.72	92.41	68.37	62.09	76.41	73.90
C2 average	99.95%	5.55	10.65	9.22	21.45	96.33	99.42	97.92	99.62	9.78	18.91	14.69	33.16
	99.90%	5.88	11.28	9.78	23.40	93.66	98.96	96.11	99.34	10.40	19.98	15.70	36.08
	99.50%	11.73	19.52	16.59	32.73	93.86	97.11	94.11	98.22	20.17	31.86	26.59	47.77
	99.00%	15.03	24.93	22.55	37.60	91.85	95.64	93.52	97.07	25.26	38.84	34.94	52.97
	95.00%	37.41	40.95	43.83	53.17	86.45	88.43	88.27	90.76	51.17	55.33	57.69	66.18
C3 average	99.95%	1.04	1.46	1.44	6.96	94.80	93.56	91.97	99.01	2.05	2.85	2.78	12.85
	99.90%	1.20	1.78	1.65	8.04	91.31	89.73	85.41	98.50	2.37	3.46	3.18	14.76
	99.50%	4.12	4.74	4.92	19.50	85.62	89.05	85.01	96.63	7.83	8.96	9.03	31.65
	99.00%	7.40	9.89	6.92	24.81	83.64	87.41	82.80	94.99	13.51	17.32	12.48	38.32
	95.00%	24.82	27.35	24.18	39.79	80.99	82.36	81.13	87.38	37.59	40.28	36.73	53.82
Overall average	99.95%	4.22	6.57	7.53	13.92	96.55	97.27	92.65	99.33	7.63	11.77	12.62	22.98
	99.90%	4.54	7.42	8.56	15.79	94.11	95.77	90.73	99.04	8.22	13.30	14.33	25.96
	99.50%	11.04	15.23	17.14	27.24	92.05	94.56	92.44	97.71	18.87	25.18	26.94	41.54
	99.00%	15.97	21.19	22.73	34.18	90.35	93.15	91.20	96.52	25.97	33.28	34.04	49.29
	95.00%	39.19	38.79	44.56	51.66	86.30	86.99	87.37	90.18	52.38	52.57	56.94	64.63

Table 3: Comparison of the areas under ROC and PR curves: averages over seven datasets and overall averages.

Dataset	AUROC				AUPR			
	Martin	PIPE2	Ding	SPRINT	Martin	PIPE2	Ding	SPRINT
C1 average	88.43	82.24	91.26	88.48	88.61	84.29	91.80	90.26
C2 average	80.50	78.18	82.41	83.23	80.36	80.32	82.67	85.67
C3 average	74.51	72.60	74.01	77.54	73.65	72.88	72.44	79.74
Overall average	81.15	77.67	82.56	83.08	80.87	79.16	82.30	85.22

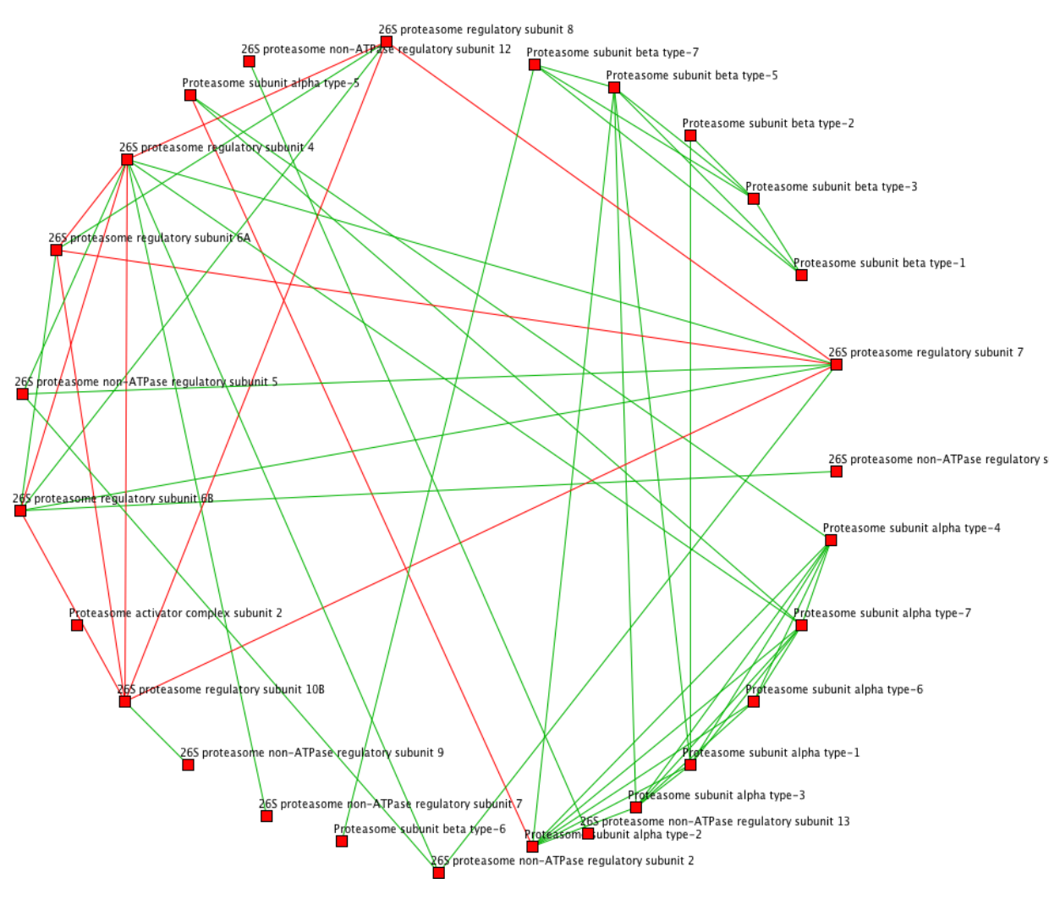

Figure 2: Example of adding predicted PPIs (red) to a known protein complex (green).

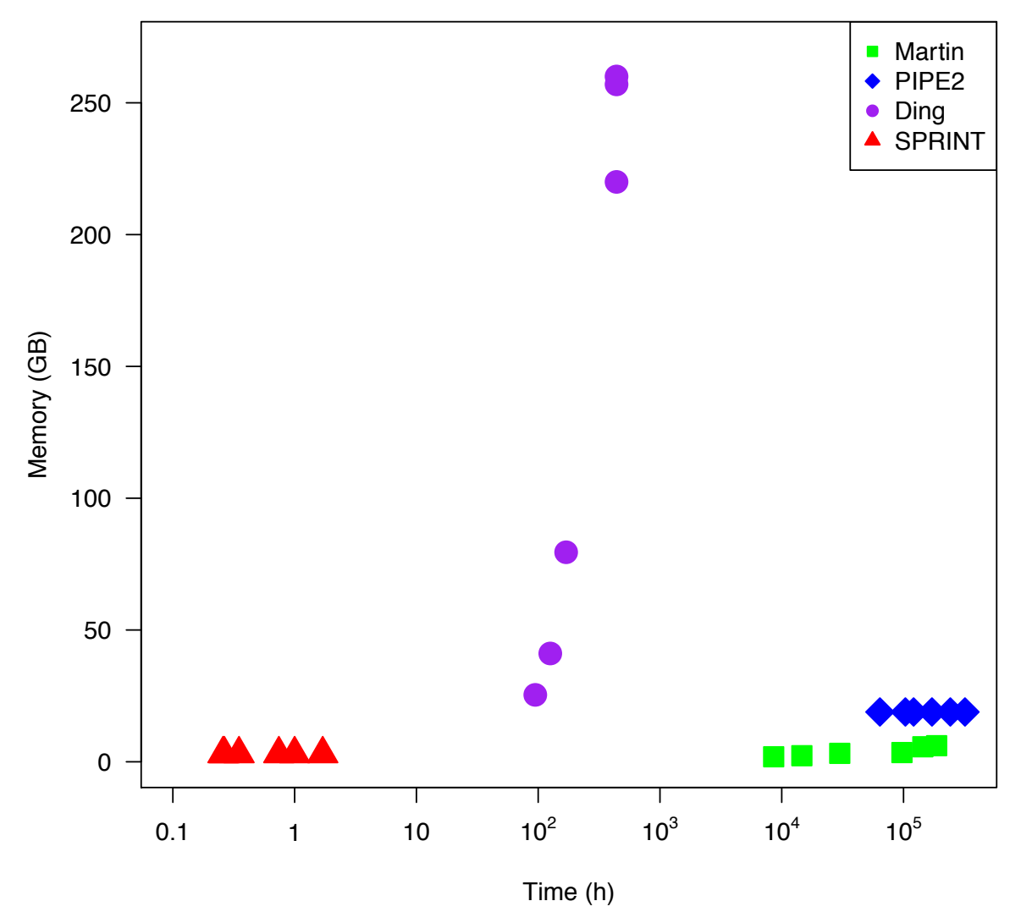

Figure 3: Predicting the entire human interactome: time and memory comparison.

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 (P_1, Q_1) is a known interaction. Each pair of blocks in P_1 and Q_1 (dashed line) increases the likelihood of interaction between proteins containg similar subsequences.

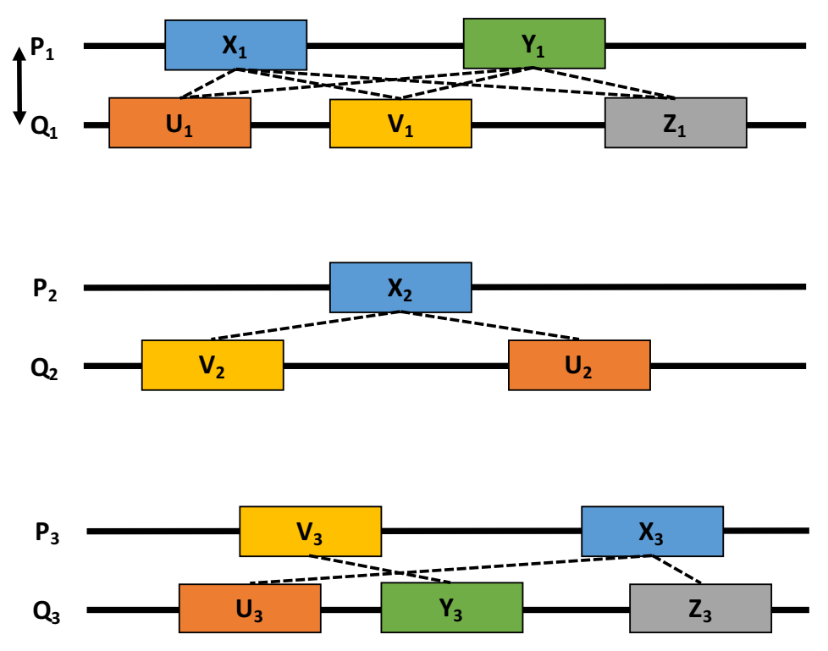

Figure 4: SPRINT idea.

Step 1: Finding similar subsequences

Highly sensitive multiple spaced seeds[4] are used for very fast and reliable computation of similar subsequences (`1` = match, $\star$ = don't care; e.g., `11****11***1`). In addition to exact hits (Figure 5(a)), approximate hits (Figure 5(b)) are essential in achieving high sensitivity. The usual k-mers are replaced by s-mers (Figure 6). Bitwise operations are heavily used for speed.

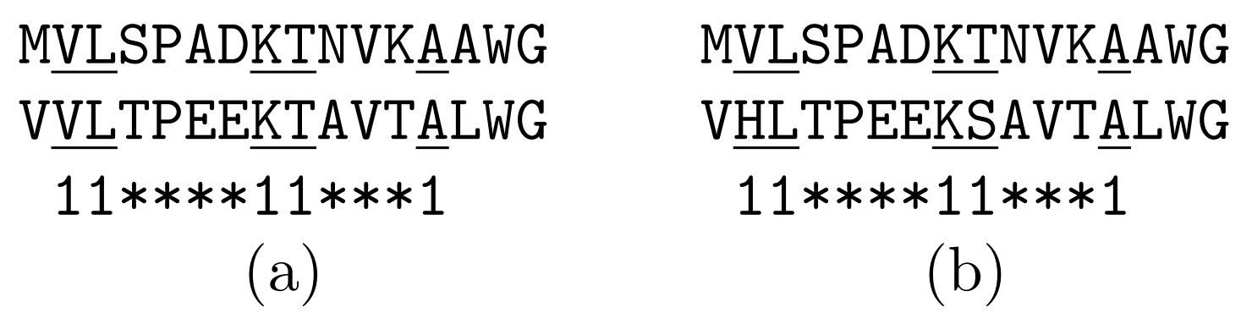

Figure 5: An exact hit (a) and an approximate hit (b).

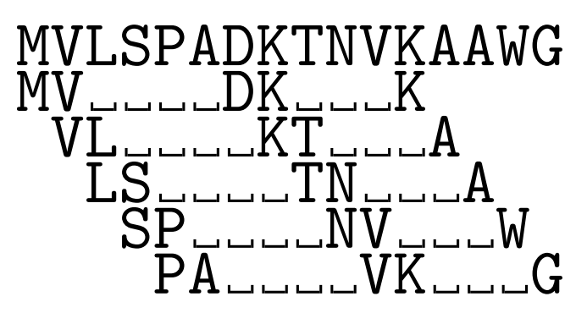

Figure 6: All s-mers of a sequence.

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 counters (5 or larger in the example) are removed (b).

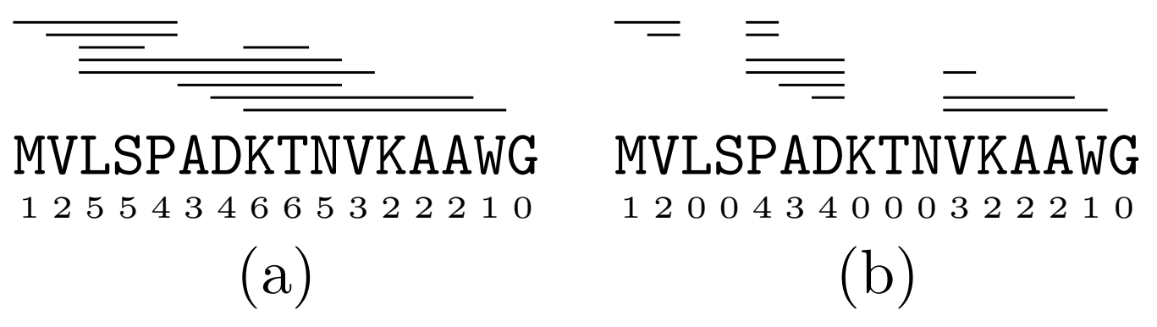

Figure 7: An example of similarities before (a) and after (b) post-processing.

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 `https://github.com/lucian-ilie/SPRINT/` and the datasets and predicted PPIs from `http://www.csd.uwo.ca/faculty/ilie/SPRINT/`.

References

[1] Park Y. Marcotte E.M. *Nature Methods*, 9(12):1134–1136, 2012.

[2] Ding Y. *et al.* *BMC Bioinformatics*, 17(1):398, 2016.

[3] Guo Y. *et al.* *Nucleic Acids Research*, 36(9):3025–3030, 2008.

[4] Ilie L. *et al.* *Bioinformatics*, 27(17):2433–2434, 2011.

[5] Martin S. *et al.* *Bioinformatics*, 21(2):218–226, 2005.

[6] Pitre S. *et al.* *Nucleic Acids Research*, 36(13):4286–4294, 2008.

[7] Shen J. *et al.* *Proceedings of the National Academy of Sciences*, 104(11):4337–4341, 2007.

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.