Working document:
structural link prediction using PPI3D dataset

1 Data and preprocessing

Starting from dataset `clustering95/human_9606/filtered_human_area_less12000.csv`

1.1 Preprocessing

- the edges considered are limited to protein-protein interactions (exclude protein-peptide)
- we exclude `homo` edges (from the protein to itself)
- we consider that `s1_seq_cluster_95` can be used as a proxy to identify a given protein

1.2 Characteristics of the PPI network

$n = 10,077$ proteins, $m = 12,957$ distinct interactions $\Rightarrow$ average degree $\overline{d} = \frac{2m}{n} \simeq 2.57$ and density $\delta = \frac{2m}{n(n-1)} \simeq 2.55 \times 10^{-4}$

For comparison:

- `STRING` (confidence ≥ 900): $n = 6,743$, $m = 67,730$
- `Heine et al.` $n = 5,457$, $m = 28,779$
- `Lit-bm-13` $n = 3,391$, $m = 4,905$
- `Lit-nb-13` $n = 5,545$, $m = 11,044$

2 Basic predictions on the PPI dataset

We implement basic predictors: `L3` and variants, and `Adamic-Adar` for comparison in an unsupervised way to the data. The related Figure is 1.

Comments:

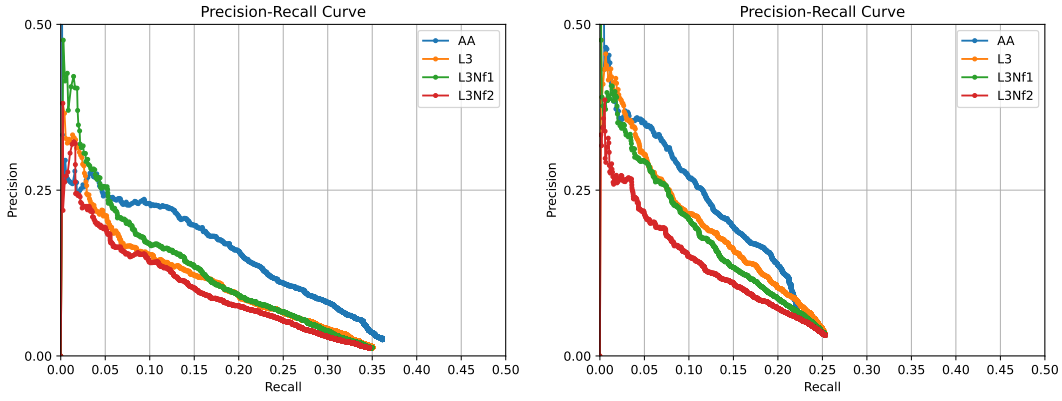

Figure 1: Left: data split (train/test) is 75-25; right is 50-50 (Yuen data split).

- the performances are not quite similar to what we find on other relatively small datasets (`Heine et al.`, `Lit-bm-13`, `Lit-nb-13`), see [YJ23]: L3 and variants are less efficient than usual, especially at larger recall ⇒ we may have performance improvement when stacking

[TODO: try systematically with pipeline:

- more datasets (all the ones in Yuen et al.)
- more data splits

→ when Raphaël has the pipeline ready]

3 Making a binding-site network

Based on the idea that the jigsaw puzzle hypothesis should work better at the scale of the binding site, we transform the network into a binding site network (BSN). For this purpose we use `s1_binding_site_cluster_95` as the identifier of the nodes.

Characteristics of the BSN (at 95): $n = 28,945$ distinct binding sites, $m = 17,714$ interactions ⇒ average degree $\overline{d} = \frac{2m}{n} \simeq 1.23$ and density $\delta = \frac{2m}{n(n-1)} \simeq 4.23 \times 10^{-5}$. It is thus very sparse, making prediction on it is hardly doable. Justas suggests to use a lower binding site similarity cluster (70 or 40), which would make it denser.

3.1 Binding-site network at clustering 70

Characteristics of the BSN (at 70): $n = 23,871$ distinct binding sites, $m = 15,770$ interactions ⇒ average degree $\overline{d} \simeq 1.32$. Now it starts being readable (Fig. 2), predictions have very low recall (the scale is different from the one in Fig. 1). Note that Adamic-Adar remains at (0,0) which means no true positive prediction.

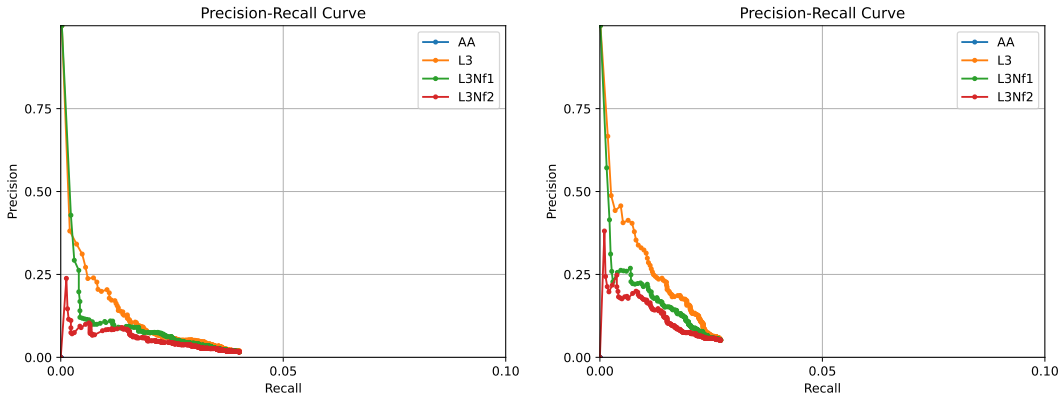

Figure 2: Prediction on binding-site network at clustering 70. Left: data split (train/test) is 75-25; right is 50-50 (Yuen data split).

3.2 Binding-site network at clustering 40

Characteristics of the BSN (at 40): $n = 22,723$ distinct binding sites, $m = 14837$ interactions ⇒ average degree $\overline{d} \simeq 1.31$. We remain at the same average degree level, and the performance are barely readable again (Fig. 3).

[LT: Maybe the poor performances that we see here are consequences of the fact that the jigsaw assumption is not correct. I this case, it may be useful to devise that a link prediction method based on the explanation that we identify works better on binding site networks.]

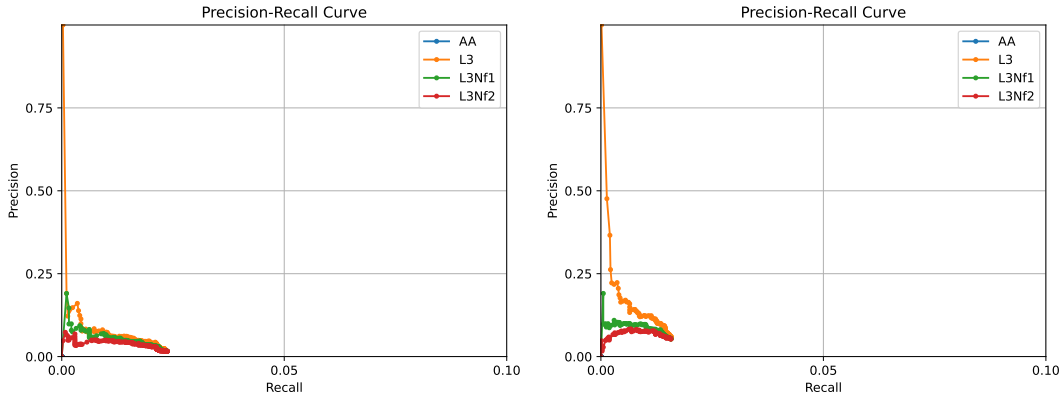

Figure 3: Prediction on binding-site network at clustering 40. Left: data split (train/test) is 75-25; right is 50-50 (Yuen data split).

4 Adding biology to the L3-prediction

4.1 Testing the jigsaw puzzle assumption

According to the puzzle assumption, if we observe the following edges: `x-u`, `u-v`, `v-y`, then `y-x` should be present in the network. In other words, if we see a path-3 pattern in the network, we should observe that this path is closed in a length 4 cycle.

We measure in the PPI network how frequent this pattern actually is. We count 147,727 path-3 patterns and 19,557 cycles of length 4. As there are 4 such paths per cycle, we can count a normalized ratio (let's call it 4-transitivity), which is the probability for a 3-path to be completed in a 4 cycle in the dataset:

$$\frac{4.\square}{\sqcap} = 0.530$$

For comparison, if edges were totally random, then it should be equal to the density $\delta \simeq 2.55 \times 10^{-4}$. This is thus very high.

We can try to define a transitivity-based prediction to see if it can challenge L3 performances. [TODO: low priority]

Also, when listing theses squares, we observe many structure like this:

```
85339 85346 85340 85351
85339 85346 85348 85351
85340 85346 85348 85351
etc.
```

which seems to indicate that many cycles are part of a larger structure, which would appear as a larger quasi-clique. It led to the assumption in subsection 4.3.

4.2 Another test of the jigsaw puzzle assumption

[EL: For testing the similarity hypothesis, let's say we have the scenario with x-u, u-v, v-y described, and x indeed interacts with y in the network, so we have a 4-cycle. We know all the binding sites from PPI3D. So we could test whether: the x-y interface is within the same cluster (clustering thresholds to define) as the u-v interface, and also the x-u interface and the v-y interface, because this is the jigsaw puzzle assumption, right? (Interfaces could be replaced here by binding site: is the u-binding site of x in the same cluster, meaning similar, to the u-binding site of v? And is the v-binding site of y similar to the v-binding site of u?]

Cycles classification. As a first attempt to do that, I suggest to classify the cycles (x,u,v,y) in this way:

- category 1 (eligible for jigsaw): for each node there is a common binding site involved in both the edges it is part of, e.g. if site A of `85339` is involved in the edge `85339 85346` then site A must be involved in edge `85351 85339`.
- category 2 (ineligible for jigsaw): at least one of the four nodes does not follow that rule

Then we refine category 1:

- category 1-1 (full jigsaw): we observe that the binding site of x can be the same as the one of v (let's note that $x \equiv v$) and $u \equiv y$
- category 1-2 (partial jigsaw): $x \equiv v$ xor $u \equiv y$
- category 1-3 (no jigsaw): not $x \equiv v$ and not $u \equiv y$

Following this classification, the experiments lead to the results in Table 1.

	threshold 95%	threshold 70%	threshold 40%
category 1	227	282	283
category 1-1	0	9	44
category 1-2	0	133	147
category 1-3	227	140	92
category 2	19330	19275	19274

Table 1: Classification of cycles according to their agreement with the jigsaw puzzle assumption, as a function of the of the binding site cluster threshold.

My observations are as follows:

1. an overwhelming majority of cycles are not even eligible to satisfy the jigsaw, because at least one node has not the same binding-site involved in the two edges involved in the cycle. [LT: Maybe it is too strong a criterion?]
2. the threshold does not change significantly the repartition between categories 1 and 2.
3. the threshold changes the repartition among category 1 cycles: at high threshold none of them satisfy the full-jigsaw assumption, while at a low threshold some of them do.

A closer look. Looking at the 44 category 1-1 cycles observed for a binding site similarity threshold of 40%, we can see that the binding site cluster `13587_0` of protein and `14270_1` are involved in 29 out of the 44 cycles. This could indicate some kind of hyper-connected binding sites

Here is the comprehensive list of 29 cycles: `(24600, 26278, 29490, 27741) (27657, 29490, 27741, 30612) (24600, 26278, 30612, 27741) (26278, 29163, 27741, 30612) (26586, 29490, 27657, 31080) (27741, 29490, 60940, 30612) (26778, 29163, 27659, 30845) (27657, 29490, 27967, 30612) (26278, 29163, 27741, 29490) (25986, 29357, 60940, 29490) (26278, 29163, 27967, 30612) (26278, 29490, 60940, 30612) (26278, 29163, 27967, 29490) (26278, 29490, 27741, 30612) (24600, 27741, 30612, 27967) (26278, 29490, 27967, 30612) (27741, 29163, 27967, 30612) (27657, 29490, 60940, 30612) (27967, 29490, 60940, 30612) (24600, 26278, 29490, 27967) (27741, 29163, 27967, 29490) (24600, 26278, 30612, 27967) (26278, 29490, 27657, 30612) (27741, 29490, 27967, 30612) (24600, 27741, 29490, 27967) (24600, 26278, 29163, 27741) (24600, 26278, 29163, 27967) (25986, 29490, 26926, 30065) (24600, 27741, 29163, 27967)`

4.3 Nature of the edges predicted by L3: superstructures?

Based on this naked-eye observation, Justas suggested that the 4 edges cyclic structures are actually part of the same "superstructure".

Direct test: we measure if the edges `y-x` are actually part of the same superstructure.

First approach. In most case, we can assume that the interactions share the same pdb identifier `pdbid`. We measure in the dataset if it is indeed the case in the cycles formerly enumerated. I did that for the 19,557 cycles, and I find:

- 2,708 cycles where all `pdbid` are identical (of the form `9cwt 9cwt 9cwt 9cwt`)
- 6,574 cycles where there are 2 different `pdbid` (of the form `9cq4 9cq4 8za6 8za6`)
- 6,535 cycles where there are 3 different `pdbid` (of the form `9cn3 3j7y 7o9m 7o9m`)
- 3,740 cycles where all `pdbid` are different (of the form `7w3m 9m2w 8usb 9e8o`)

In my opinion, it rather goes in the sense of Justas' assumption that in most cases, we have superstructures containing 4-cycles and that's what L3 would mostly predict.

Also, it is an underestimate because the database is made in a way that might lead to a different `pdbid` even if the interactions are involved in the same structure. The reason is that the `pdbid` selected is one of the "best" one, and if the PPI is involved in several structures, it might not be the one corresponding to the superstructure that we are looking for.

Improved search for `pdbid`. Now, I list all the `pdbid` involving the edges according to the original dataset, and test for each cycle which combination of `pdbid` minimizes the diversity of different `pdbid` in the cycles. The 19,557 cycles are now distributed as such:

- 4,648 cycles where all `pdbid` are identical (of the form `9cwt 9cwt 9cwt 9cwt`)
- 9,986 cycles where there are 2 different `pdbid` (of the form `9cq4 9cq4 8za6 8za6`)
- 3,139 cycles where there are 3 different `pdbid` (of the form `9cn3 3j7y 7o9m 7o9m`)
- 1,784 cycles where all `pdbid` are different (of the form `7w3m 9m2w 8usb 9e8o`)

An important fraction of cycles involves a superstructure.

[EL: Most of these 4-cycles seem to correspond to superstructure ⇒ is this a feature specific to the PDB or should we expect that this could also play a role in the success of L3 and variants on STRING? What would be a good metric to detect superstructures with missing links in a PPI network?] [LT: Actually, I don't know if it is possible to have the info of the superstructure with STRING, is it?]

[EL: In such cases, I guess that the assumption that

- the u-binding site of x is similar to the u-binding site of v, and to the y-binding site of x
- the v-binding site of u is similar to the v-binding site of y, and to the x-binding site of y

Does not hold since each protein is in contact with multiple partners at the same time.][LT: That's true, so if it is the main reason for L3 predictions, it should not work correctly with a binding site network.]

[EL: I would guess that if you consider larger sets of 3-paths that share common nodes (variant of L3), you would have fewer of these superstructures?]

[EL: Maybe we should focus on proteins that are already known to interact with many partners through the same binding site, like calmodulin: `https://string-db.org/cgi/network?taskId=bsrEKcvTXRVB&sessionId=bpndJtPMXdER` List of PDB codes where it appears: `https://www.uniprot.org/uniprotkb/P0DP23/entry`]

4.4 Nature of the edges predicted by L3: interactions involving unstructured parts of the proteins?

Another possibility could be that the edges predicted by L3 correspond to interactions involving unstructured parts of the protein.

One possibility could be to use AlphaFold to give an estimate of the fraction of the protein that is unstructured to see if it is the case.

[LT: for the moment, I don't know how to address that properly.]

[EL: This would in any case invalidate the jigsaw puzzle assumption.]

[TODO: low priority]

5 Complementarity measures and L3 prediction

I want to know if the L3 predicted edges (or any measure predicted edges actually) correspond to more complementary proteins.

This suppose to define some notion of complementarity, itself based on the notion of similarity. We could use `s1_seq_cluster_70` or `s1_seq_cluster_40` as an intermediate to compute the notion of similarity between proteins.

[LT: For the moment I did not really address this topic.]

[EL: Complementarity is tricky because this brings us back to the classical docking problem.]

[TODO: low priority]

References

[YJ23] Ho Yin Yuen and Jesper Jansson. Normalized l3-based link prediction in protein–protein interaction networks. *BMC bioinformatics*, 24(1):59, 2023.