

Humboldt University Team Finds Little Overlap in Eight Human Interaction Maps

April 13, 2007



Matthias Futschik,
Institute for
Theoretical Biology,
Humboldt University

The number of publicly available resources for human protein-protein interaction data is on the rise, but a recent study led by researchers at the Institute for Theoretical Biology at Germany's Humboldt University indicates that these resources have very few interactions in common.

The study, published in the online version of [Bioinformatics](#) in January, compared eight publicly available human interaction maps of three different types — manually curated, computationally predicted based on interactions between orthologous proteins in other organisms, and large-scale yeast two-hybrid scans.

The researchers note in the paper that they were surprised at how little overlap they found between these resources: Out of a total of 10,769 proteins, only 10 were found in all eight maps; and out of 57,095 interactions, none was found in six or more maps.

As part of the project, the Humboldt University team developed its own integrated human protein interaction database called the Unified Human Interactome, or [UniHI](#), to provide access to all available human protein interaction information from one entry point.

BioInform spoke with Matthias Futschik, the lead author on the paper, this week to discuss these findings and their potential impact on biological research, and to get additional information about UniHI.

Considering that the initial comparisons between yeast protein-protein interaction maps in 2002 generated so much interest, it seems odd that no one has compared human interaction maps prior to this. Why do you suppose this is the case, and what drove you to undertake the challenge?

In contrast to yeast, the human interaction networks haven't been around for such a long time. The literature-based ones were developed in 2000 or 2001, and the orthology-based ones, which are purely computational, started in 2003, but four of the eight networks we compared were only available in 2005.

We actually started at the end of 2005 and it originated from a collaboration with Erich Wanker [of the Max Delbrück Center for Molecular Medicine], who is a coauthor on one of the interaction networks. They did yeast two-hybrid screens and they published in *Cell* [A human protein-protein interaction network: a resource for annotating the proteome. *Cell*. 2005 Sep 23;122(6):957-68] — the first human interaction network based on yeast two-hybrid screens.

So because we were in a collaboration, we started to compare it to other available networks, and at this time there was no second yeast two-hybrid, so we compared it with [the Human Protein Reference Database], and then when the other yeast two-hybrid came out, the one by [Marc] Vidal's group [Towards a proteome-scale map of the human protein-protein interaction network. *Nature*. 2005 Oct 20;437(7062):1173-8], we were surprised that not many interactions were in these three networks.

So this was the motivation for why we started to systematically examine the coherency and the concurrency between and within interaction networks. We wanted to see if our first results were typical for the current state of interaction networks, or whether they were because the yeast two-hybrid maps were maybe outliers.

But it turned out that the low number of interactions in common between networks is quite a common feature in the current stage of the human interactome.

You said you were surprised that there were so few interactions in common, but wasn't that in line with what was seen with yeast? Or have there been improvements in the yeast two-hybrid methodology over the past few years that led you to believe the results would be better in human?

Actually, we thought so, because ... these techniques have improved over the last few years, so we assumed that you would catch more interactions that are known in other interaction networks.

But that's obviously not the case.

No, and there are probably many reasons. One reason is that maybe in contrast to yeast, for human proteins the interactome is probably more dynamic. So you won't find some protein modifications when you're using only yeast two-hybrid interaction screens.

You mentioned that you looked at the coherency and concurrency of these networks for this analysis. Can you explain what you mean by those terms and why they're significant for this study?

Concurrency is just how many interactions of one network that you find in another network. So it's just the overlap. The total overlap was quite small, but we saw some quite considerable tendencies. One important finding was that networks generated by the same approach have got a larger overlap. On the positive side, this means they are probably reproducible because if they weren't there would be a random overlap.

On the other side, they have got tendencies, or internal biases, in them. So we also checked if there is an enrichment for proteins of certain functions. One example was that there is a large enrichment in literature-based networks for signal transducers or regulatory proteins. So this means they are well studied in these networks.

With coherency, we took several approaches and one approach is, based on the knowledge we have about proteins, one can assume that proteins of the same function act together. So we checked if this is true for the interaction networks, and you see that they are all of them coherent to a certain degree. Of course, the literature-based ones perform better in this test, but then you must also say that the knowledge about function is derived from the literature, too. So it's not a truly independent benchmark test because it's maybe derived from the same publications. So it's a catch-22 a little bit.

So to avoid this, we also used expression data from the Gene Atlas. This is a large expression data set with different human tissues, and we checked if interacting proteins are co-expressed, because people have found that in yeast, frequently interacting proteins are co-expressed. So we checked this too, and you get quite high co-expression relative to random, and co-expression to literature-based ones and orthology-based ones, and to a lesser degree for yeast two-hybrid networks.

But again, with each of these tests, you can say on the one side it catches a feature, but on the other side, there are a lot of interactions where we know they are transient, so we know they don't need to be co-expressed. Often this co-expression only exists in the stable complexes.

Would you consider any of the three types of networks to be more or less reliable than the others? Or is it just a question of knowing what the limits of each of them are?

On the one side, one has to ask, 'What's reliable?' because in contrast to the DNA in the genome, the interactome is not stable. So under different conditions, different proteins, it heavily depends on modifications, so when you look at these interaction networks, and you see these hairballs, this is a projection of many, many conditions at the moment. That's one of the major problems I think in the field, that we have to define more conditions. So some proteins may interact under one condition — in, for example, a liver cell — but under other conditions — in a nerve cell — they would hardly interact or not even be expressed.

So the reliability depends on the conditions. Most people would say the literature-based ones are more reliable, and people use them as the gold standard. And I think that maybe they are more reliable than the two others, but they definitely have got a high false positive rate, too, in the sense that there is a heavy inspection bias. And several papers will come out, by us and by other groups, that probably will point this out — that if you are really looking hard for some interactions between some proteins, then you will find them.

So in a way, I think it's important to see the limitations of these approaches, and to know what biases and what pitfalls they have.

What does this mean for researchers who are using this data? As you mention in the paper, one goal for these networks is to use them as frameworks for modeling and simulation in computational systems biology, so what effect would the current state of this data have on these efforts?

I think it means that they first have to say where this data comes from ... [and] to see if such a network is actually expressed at the same time under the same conditions. So just taking out a network and trying to do modeling with it, that won't be very reliable.

Therefore, for example, with UniHI, our database, we're trying to provide one more resource. Our philosophy is not to give a final network because this will depend on the conditions. If you're looking at a certain cell type it might look totally different from another cell type, so we want to give researchers a tool so that they can say, 'I'd like to search for protein interaction networks that are specific to the brain because I'm interested in neurodegenerative diseases.'

Our aim is to provide really more of a dynamic network and to give integrative tools for researchers to get out of this mass of interaction data the most suitable data for the task they want to solve.

So UniHI provides a sort of filtering capability to create subnetworks based on particular cell types or experimental conditions?

Yes. And our philosophy is to link up to different databases, so UniHI should be an entry gate to the human interactome. We don't say, 'Just use the literature ones or just use the yeast two-hybrid ones because they're experimentally verified and essentially tested.' We say, 'OK, here are the links.' We give them a lot of filtering possibilities, so you can straight away say, 'I don't use the orthology-based ones because they are computationally based, and I only want to have interactions that are experimentally tested.'

We also want to give real-life information about where the genes are corresponding to the protein expressed. So in some tissues, if you don't find expression you wouldn't expect the interaction of the corresponding proteins. And we also want to give filtering functions that people can say, 'I only want to have interactions that are reproduced by two experiments or by two different experimental techniques.'

On the one side, there is of course a [requirement] that the user has some insight as to what he's looking for. So we're not presenting something that's just plug and play. What we want to present is, for a researcher who is competent in the field of interaction networks and interaction network modeling, a tool that is as flexible and as powerful as possible, but also narrows down this large interaction space to some manageable format that they can use for systems biology.

So this is more of a way to manage the available information from different resources rather than a monolithic resource of all of those interactions in one place.

We don't think there is a static, ultimate interaction network, but there are lots of interaction networks out there and we want to link them and just give researchers the freedom to look for their specific ones based on other information.

The idea of UniHI was not to download everything and put it in one place, but to provide links to all of the interactions. We put together all the networks, but we provide links to the original databases so that people can go and find more information and more annotation about specific interactions.

So it's an entry gate. ... You have to find your own way, but we try to give as much information as possible.