BIOINFORMATICS EDITORIAL Vol. 22 no. 24 2006, pages 2973–2974
doi:10.1093/bioinformatics/btl579

Bioinformatics in the human interactome project

'*In the early days of the Human Interactome Project, a meeting was organized...*'. Perhaps, a few years from now, newspapers will describe in those terms how straightforward it was to plan the large-scale mapping of protein interactions in human and other model organisms. Scientists attending the second Cold Spring Harbor Laboratory/Wellcome Trust symposium on 'Interactome Networks'[1] know well that things are not that easy. Important scientific, technical and sociological issues remain before the 'Human Interactome Project' can be considered on its way. But things are definitely moving.

At the meeting, Marc Vidal proposed some concrete goals for such a project: 'To produce hundreds of different sets of cloned ORFs (ORFeomes) and 100 million interactions with a 1–5% false positive rate. To add directionality and signs to the interactions (i.e. activation or inhibition), and to study the variation of the interactions associated with diseases'. Nonetheless, the community still has much to decide. It will still have to agree on these goals, to subdivide the project into recognizable milestones, to set a time line for achieving these milestones, to associate cost to each of the operations, and perhaps most importantly, to obtain funding for such an ambitious endeavor. But there is little doubt that having clear goals will help strengthen the ties between researchers in this already very active community, as well as to engage new partners and grant agencies.

As with the Human Genome Project, bioinformatics and computational biology will be of profound importance to any protein interaction mapping effort. Following the presentations during the meeting (for a recent review see Sharan and Ideker, 2006) an early and essential bioinformatics task will be the creation of database standards (i.e. the IMEX interaction database standard and the emerging Biopax bio-pathways standard) and analysis/visualization platforms, such as the one provided by the Cytoscape project. It was also interesting to realize the progress that has already been made in the analysis of the structure, function and properties of protein interaction networks (as well as gene control and metabolic networks), even while the number of reliable datasets is still small. It is also rewarding to see also how the first large-scale simulations based on protein interaction data are becoming a reality.

These efforts in analysis and simulation are proceeding in parallel with those dedicated to the prediction of new interactions, modules, motifs and functional properties (phenotypes, diseases and others), in most cases by integrating complementary sources of information on functional and structural interactions. All this domain of emerging 'Network Biology' offers a direct connection between computational and the experimental analysis. In this respect, two questions emerge from the meeting as critical for the future: (1) the development of methods able to distinguish physical from functional interactions (and/or different types of physical interaction) and (2) the mapping of the details of the physical interactions (i.e. interacting residues) and other information that is necessary for the interpretation of the variation data (SNPs) and for experimental manipulation of interaction networks.

Finally, an interesting controversy arose during the meeting that might have consequences for our bioinformatics community. Some argue that it will be more effective to concentrate all efforts into scale-up of the experimental proteomics technology, postponing the bioinformatics analysis to a second phase once the underlying data are fully (or at least mostly) complete. On the contrary, we think that, it is essential to continuously support the development of the methods that will be required for the interpretation of the Human Interactome Project, including network alignments, annotation, analysis and others. The analogy with the Human Genome Project can be useful here. In that case even if the basic alignment techniques were ready since the 70's when the genome sequencing emerged basic bioinformatic technologies were not available (c.f. just remember the challenging analysis of the first bacterial genome in 1995 (Casari *et al*., 1995), or the struggle to assemble and represent the first draft of the human genome (Istrail *et al*., 2004). We are convinced that by pushing in parallel experimental and computational developments we can prepare in a more effective way the future of this area of research. Indeed, much of the current interest in large-scale proteomics is related with the impact that the early computational analysis of the first (and imperfect) datasets have had (i.e. the first 'scale free' and 'motif discovery' papers of Barabasi (Jeong *et al*., 2000) and Alon (Shen-Orr *et al*., 2002) teams have captured the imagination of biologist, physicists and theoreticians like few other problems in molecular biology have). Moreover, integrative and computational approaches have already been indispensable for assessing data quality and scoring confidence in specific interactions as well as whole interaction datasets. Finally, at a practical level, what biologists see as a result of large-scale proteomics are computational representations based on the data provided by databases. Therefore, a successful Human Proteome Project depends intimately on ongoing developments in bioinformatics, as they proceed in parallel with the large-scale experiments.

[1]'Interactome Networks' Joint Cold Spring Harbor Laboratory/Wellcome Trust Conference, August 30–September 3, 2006, Hinxton UK. Organized by Ewan Birney, EBI; Anne-Claude Gavin, EMBL; Marc Vidal, Dana-Farber Cancer Institute, Harvard Medical School, USA.

REFERENCES

Casari,G. *et al*. (1995) Challenging times for bioinformatics. *Nature*, **376**, 647–648.

Istrail,S. *et al*. (2004) Whole-genome shotgun assembly and comparison of human genome assemblies. *Proc. Natl Acad. Sci. USA*, **101**, 1916–1921.

Jeong,H. *et al*. (2000) The large-scale organization of metabolic networks. *Nature*, **407**, 651–654.

Sharan,R. and Ideker,T. (2006) Modeling cellular machinery through biological network comparison. *Nat. Biotechnol*., **24**, 427–433.

Shen-Orr,S.S. *et al*. (2002) Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat. Genet*., **31**, 64–68.

Trey Ideker[1]
Alfonso Valencia[2]

[1]Department of Bioengineering, University of California, San Diego (UCSD), USA

[2]Structural and Computational Biology Programme, Spanish National Cancer Research Centre (CNIO), Melchor Fernandez Almagro, 3. E-28029 Madrid, Spain