Integrating biological process modelling with gene expression data and ontologies for functional genomics

(Position paper)

Liviu Badea, Doina Tilivea

[1] AI Lab, National Institute for Research and Development in Informatics
8-10 Averescu Blvd., Bucharest, Romania
badea@ici.ro

In the current post-genomic era, various aspects of gene function are being uncovered by a large number of experiments producing huge amounts of heterogeneous data at an accelerating pace. Putting all this data together, while taking into account existing knowledge has become a pressing need for developing environments able to explore and simulate biological entities at a *system level*.

We argue for the need to create a *common bioinformatic framework for modelling biological processes* by a non-trivial *integration* of various complementary functional genomics data and knowledge with the goals of representing and simulating the relevant biological networks and pathways, discovering targets for drugs and diagnostics, as well as determining the molecular mechanisms of diseases from gene expression data. Such a synergetic use of the available data will allow partly replacing certain costly, or even impractical biological experiments by "in silico" simulations of biological processes, as well as enable new in-depth and large-scale experiments.

There are currently at least three types of extremely valuable resources, which are currently not used at their full potential:

- *gene expression data* (e.g. from microarray experiments)
- knowledge about *networks and pathways* (such as metabolic, genetic control, signalling pathways and protein interaction maps)
- *ontologies* (such as the Gene Ontology).

For example, gene expression data are extremely useful for understanding gene function at a global level, but they are typically used *without taking the relevant network and pathway knowledge into account*. (This is due not only to the incompleteness of pathway databases, but also their limited representation and interoperability.)

Also, a lot of effort has been put into manual construction and annotation of pathways – this valuable knowledge should be used in as many contexts as possible. Unfortunately however, the modelling languages used by various annotation efforts are slightly different in terms of expressiveness, making the fusion of knowledge from such different pathway databases a difficult knowledge modelling problem.

Modelling language, ontologies

The main goals of system biology are related to modelling biological entities at a system (holistic) level for various purposes: analysis, simulation, prediction, etc (listed in increasing order of complexity). Thus, while analysis does not necessarily require complete models of the systems involved, simulation and especially prediction are not feasible without complete knowledge (at least at a certain level). For example, the

structure of a genetic network involved in organism development may not be enough for simulation or prediction without detailed knowledge of various reaction parameters, which might be hard to obtain for all reactions.

Thus, one of the most important requirements of this field is to allow an as detailed as possible representation of all relevant aspects of the biological processes to be modelled. Still, since most existing knowledge is either very detailed, but covering only a very few biological processes, or much less detailed, but having a wide coverage, we argue that a very expressive modelling language will not do, since it will not be able to deal with the more sketchy (less detailed) knowledge that is available today. Still, the knowledge of the structure of the system will allow certain qualitative conclusions to be drawn, even in the absence of numerical parameters. Being able to exploit various heterogeneous resources for reasoning about biological systems at various levels of detail seems to be the major challenge in the field.

The modelling language thus needs to be able to describe both qualitative (e.g. structural) and quantitative aspects of a model at various levels of detail, in order to allow the integrated use of practically all existing biological knowledge, ranging from expert-curated pathway databases (such as KEGG [8], TRANSFAC/TRANSPATH [13], CSNDB) to high-throughput but less detailed experimental data such as protein-interaction data, or even putative computationally-derived annotations. The language should allow a *gradual transition from less detailed qualitative knowledge to very detailed quantitative knowledge about biological mechanisms* and *the use of such partial models during reasoning at all intermediate stages*.

Since the models should also be processable by computer programs (and not only by human biologists, as it is still the case today[1]), the modelling language will have to have a precisely defined formal semantics, that would allow the correct interoperation of the various software modules using them.

Of course, all of these features require much more sophisticated reasoning tools. While for a *uniform* modelling language (even a very expressive one, based for example on partial differential equations), reasoning is relatively easy using existing tools, reasoning in a heterogeneous modelling language allowing descriptions at various levels of detail is highly non-trivial and should rely on *open* architectures (open both from a technical and a conceptual point of view). *Technically*, the architecture should allow the integration of various software modules (e.g. PDE solvers, numerical packages, symbolic reasoning tools such as abductive reasoners, constraint satisfaction modules, process simulation and analysis tools, etc.). *Conceptually*, there should exist a unified model able to view the various types of knowledge in a uniform framework. A potential candidate for such an *open* modelling unified architecture could be a high-level constraint reasoning environment such as a Constraint Logic Programming (CLP) system allowing the declarative implementation of constraint solvers using Constraint Handling Rules (CHR) [3].

Developing adequate representations for genes, proteins, networks, pathways, etc. is crucial for developing an integrated framework for molecular biology and genetics data. Current representations are rather fragmentary – a much tighter integration of the

[1] Many pathway databases (e.g. KEGG) are currently more oriented towards a human user interface, rather than a computer processable one.

various representations is required. Such representations have to refer a common vocabulary of terms (in molecular biology / genetics), such as the *Gene Ontology* [4].

Since the resources in this field are distributed and currently accessible via Web-based interfaces, it is important to make their content accessible in a "semantic Web" format (e.g. using newly proposed standards such as DAML+OIL [5]). Such enhanced representations allow not just *expressive* constructs with a formally defined semantics, but also automated reasoning about them (without such inference services, the representations are useless w.r.t. automatic processing, which is essential in a field involving huge amounts of data and knowledge).

We target the following aspects:

- **Modelling various types of biological networks and pathways** (metabolic, genetic control, signalling networks) in a unified framework that should also allow their simulation as well as automated reasoning. In our mind, it is important to devise a representation formalism for biological pathways that is not only very expressive, but also usable by sophisticated reasoning services (such as matching subnetworks by complex logical descriptions of molecular disruptions of a target disease).
- **Devising ontologies** more sophisticated than e.g. the Gene Ontology (which is just a hierarchy of molecular biology/genetics terms), possibly with domain-specific constructs, having a limited scope. An essential part still missing in all existing ontologies is the information about networks and pathways, which are essential for the new emerging field of Systems Biology. For example, a 'molecular interaction' might be not just a vocabulary term, but also a complex object (possibly similar to a transition in a Petri net) with associated components (in this case substrates/products) as well as reasoning components (that can be invoked to reason about such interactions).

The ever-growing amount of experimental data in molecular biology and genetics requires its automated analysis, by employing sophisticated knowledge discovery tools. In [1] we used an Inductive Logic Programming (ILP) learner to induce functional discrimination rules between genes studied using microarrays and found to be differentially expressed in three recently discovered subtypes of adenocarcinoma of the lung. The discrimination rules involve functional annotations from the Proteome HumanPSD database in terms of the Gene Ontology (GO), whose hierarchical structure is essential for this task.

It is encouraging that the discriminations obtained are biologically sensible – this heavily relies on the GO and the HumanPSD annotations. But this also automatically prompts the question of whether more sophisticated knowledge representation formalisms, such as Description Logics (DL) might allow even more precise functional distinctions to be made.

A DL may allow an "on-the-fly" construction of concepts, rather than relying on a fixed hierarchy. Thus, we wouldn't need to explicitly record in the ontology all generalizations of existing concepts. For example, the current GO contains not just specific concepts like '*cyclin-dependent protein kinase inhibitor*' or '*transmembrane receptor protein tyrosine kinase activator*', but also their generalization '*kinase regulator*'. On the other hand, a DL may take advantage of the *intrinsic composite nature* of the concepts above and represent them as $\exists inhibits.CDK$ and $\exists activates.TRPTK$. Their generalization need not be explicitly represented, since it can be computed by taking the *least general generalization* ("least common subsumer" in DL terminology) $\exists regu-$

lates.kinase of the two concepts above. Using a DL would also simplify the update and maintainance of the ontology by not having to explicitly specify all the inheritance relationships of a new concept, as well as by providing automated consistency checking tools. Another useful extension of GO would involve integrating it with metabolic, regulatory and cell signalling pathway databases (which would allow more precise causal reasoning – for example, determining possible primary causes for complex genetic disruption profiles).

Such more sophisticated representation formalisms for pathways and genes/proteins would allow mapping gene expression data onto the pathways (thereby generating pathway *activations*), which could be used in various abductive and inductive reasoning mechanisms involving:

- disease recognition (by matching pathway activations to logical descriptions of the molecular mechanisms of diseases)
- discovery of disease mechanisms by mining pathway activations.

Knowledge discovery and machine learning

A combined use of microarray gene expression data, pathways and functional annotations (e.g. in terms of the Gene Ontology) in a common framework enables not just the *interpretation* of experiments in a detailed biological context, but also the *discovery* of various types functional knowledge. Sophisticated machine learning algorithms able to deal with background knowledge (such as currently known pathways) are needed to achieve this. Already the most basic background knowledge on functional annotations, which involves *hierarchies of concepts* (as in GO, where the main type of relational information is in the form of *inheritance* relationships), is not directly treatable by propositional learners like C4.5, but could be dealt with our approach from [1]. We argue the need to go beyond attribute-value learners, i.e. towards relational learners [7], which are able to deal with structured representations and sophisticated background knowledge. Still most existing learners are either employing covering-based methods (which are inappropriate in this context), or are learning a large number of association rules, without any simple means of selecting the most relevant ones. A precise measure of interestingness of induced rules w.r.t. a given background knowledge is needed.

Pathway reconstruction

A modelling language for system biology – even a very expressive one – is useless without the ability of aquiring knowledge about significant portions of the biological processes of interest. Knowledge aquisition is particularly difficult in biology not only due to the sheer number of entities involved, but also due to their heterogeneity. We have already advocated the need for reusing existing knowledge in whatever form it might currently exist. Sometimes, however, this is not enough and large-scale manual aquisition is not only very expensive, time-consuming, but also prone to errors.

An alternative to manual knowledge aquisition in this domain could be the *automated reconstruction of biological pathways*. There are currently several types of knowledge relevant to the reconstruction (refinement) of biological networks and pathways:

- partial knowledge about pathways (from pathway databases, textbooks, literature)
- gene expression data (e.g. from microarrays experiments)
- functional annotations (e.g. in terms of the Gene Ontology from the Proteome databases).

There are also many approaches and algorithms for reconstructing pathways, but they have several important limitations:

- they typically deal with just the raw expression data (without taking into account partial knowledge about pathways or functional annotations that are already available)
- they typically require an impractically large number of knockout (or other genetic modification) experiments in order to be able to learn a *complete* network [9]. We would like to have a more incremental (i.e. an anytime) pathway reconstruction algorithm that is able to refine partial pathways in an incremental fashion.

Existing gene expression compendia (such as the Rosetta compendium for yeast [6]) contain data from a large number of microarray experiments (including knockout experiments) involving practically all yeast genes. We argue that the preliminary analysis attempts (involving most clustering of genes and expression profiles [eg. 6]) only scratch the surface of the knowledge hidden in such compendia. The most important and challenging task involves *extracting causal influence information* from such data.

While existing data mining and machine learning approaches (such as association rules learners) extract mainly shallow associations (correlations) present in data, we are aiming at discovering the true *causal structure* of the networks of interest. (Of course, for a limited amount of experimental data, only partial knowledge about the causal structure may be inferrable. We insist on the need to be able to express and refine such partial models.) The probabilistic nature of many biological processes (as well as the unavoidable noise present for example in microarray data) requires the use of probabilistic models, such as Bayes nets. Superficially, our problem resembles that of Bayes net *structure learning* – an already very difficult research problem. Unfortunately, existing Bayes net structure learners cannot be directly employed in this domain, due to certain specificities of biological networks:

1. Biological networks are *cyclic*, may have *latent (hidden) variables* (i.e. variables not present in the measured expression data) and may be prone to *selection bias.* (On the other hand, Bayes net structure learners deal only with *acyclic* networks[2], and only few are able to deal with latent variables and selection bias).
2. Since Bayes net *structures* are not Bayesian themselves (additional experimental data may change the most likely structure significantly), we need to *infer only the features of the network that are fully justified by the data* (rather than a full Bayes net in which certain edges and/or edge orientations are not fully justified by the data, but rather represent the best scoring structure). The most popular structure learning algorithms (the *scoring-based* ones) are thus inapplicable in our setting, while none of the existing *constraint-based* algorithms (e.g. IC [10], PC, FCI [11]) covers all the requirements from (1) above.

[2] The CCD algorithm of [12] is able to deal with cyclic structures, but not with latent variables. It is also limited to linear dependencies.

3. Although the constraint-based algorithms are able to infer causal structures from purely *observational* data, upgrading them to deal with a combination of observational and *experimental* data is not straight-forward and hasn't been solved yet. (The main problem seems to be related to the relatively small sub-populations of compendium samples for each perturbation, which doesn't allow reliable independence tests.) On the other hand, scoring-based methods are easily able to deal with this problem [2], but they are unapplicable due to the problems described at point (2) above.
4. Taking into account existing *background knowledge* (in the form of partial networks) is also extremely important, especially due to the very large number of variables (genes) involved in such causal inference experiments.

We are currently developing a constraint-based causal structure learner addressing all of the above problems. Since the first phase of the algorithm is very similar to clustering methods that are very popular in this domain, we may be able to compare the deeper models constructed by such a causal learner with the more shallow clusters reported in the literature. Our preliminary experiments currently show the computational feasability of our approach, as we are currently able to deal with networks of reasonably large size (e.g. 800 genes).

In conclusion, we argue for a modelling environment for system biology that is not only very expressive, but also allows a non-trivial combination of various data and knowledge sources *at various levels of detail* and supports the *automated reasoning* about the various aspects of biological function. We also advocate the potential utility of causal structure learners for obtaining at least partial drafts of biological networks from expression data.

References

1. Badea, L.: Functional discrimination of gene expression patterns in terms of the Gene Ontology, Proc. of the Pacific Symposium on Biocomputing PSB-2003.
2. Cooper, G.F. and C. Yoo, Causal discovery from a mixture of experimental and observational data, Proceedings of the Conference on Uncertainty in Artificial Intelligence (1999) 116-125.
3. Fruewirth T. Theory and Practice of Constraint Handling Rules, JLP 37:95-138, 1998.
4. Gene Ontology: tool for the unification of biology. Nature Genet. 25: 25-29, 2000.
5. Ian Horrocks et al. Reviewing the design of DAML+OIL: An ontology language for the semantic web. AAAI 2002 http://www.daml.org
6. Hughes et al.Functional discovery via a compendium of expression profiles.Cell 102(1):109-26, 2000
7. Nienhuys-Cheng, S.H., R. de Wolf. Foundations of Inductive Logic Programming, Springer, 1997.
8. Ogata, H., Goto, S., Fujibuchi, W., Kanehisa, M.: Computation with the KEGG patway database, BioSystems, 47, 119-128(1998), http://www.genome.ad.jp/kegg/kegg2.html
9. Pe'er, D., Regev, A., Elidan, G., Friedman, N.: Inferring Subnetworks from Perturbed Expression Profiles, Bioinformatics, Vol.1, no.1 2001, pages1-9.
10. Pearl, J., Verma, T.S.: "A theory of inferred causation", Proc. KR-91, 441-452.
11. Spirtes, P., Meek, C., and Richardson, T. Causal inference in the presence of latent variables and selection bias. In Proceedings of UAI-95, pp. 499-506.
12. Spirtes, P.: `Directed cyclic graphical representation of feedback models', Proceedings of the 11th Conference on Uncertainty in Articial Intelligence, Montreal QU: Morgan Kaufmann, pages 491-498.
13. Wingender, E.:The TRANSFAC System on Gene Regulation, Trends in Glycoscience and Glycotechnology 12, 255-264 (2000), http://transfac.gbf.de/TRANSFAC/