8th Protégé Conference, 2005/07/18-21, Madrid, Spain.
Presentation and poster abstract.

Exploration of the Human Nuclear Receptor Dimerization Network with Protégé – A Case Study for Biomolecular Interaction Data Management and Analysis

Gregory Amoutzias[1] & Elgar Pichler[2,*]

[1] Manchester University, Manchester, UK
[2] AstraZeneca Research Center Boston, Waltham, MA, USA
* To whom correspondence should be addressed.

Introduction
We have integrated phylogenetic, protein-protein interaction, genetic interaction, and gene expression data to provide a comprehensive and up-to-date description of the topology and properties of the nuclear receptor (NR) dimerization network in humans.
We classified and described the different NR protein and association entities in OWL and we used the Protégé platform as a data integration and data management platform, as well as a visualization and data analysis tool that helped us in the derivation of overall regulatory motifs for the NR networks.
The assembled data represent the most complete data set on the NR dimerization network currently available.
The dynamic view on these data provided by Protégé showcases how biological interaction data can be represented in a more informative way than is possible with standard interaction databases.
Analysis of the associations between members of the NR family reveal a hub-like network with a higher degree of connectivity than anticipated and points to several regulatory motifs on the level of physical as well as genetic interactions.

Nuclear receptors – characteristics, evolution, and dimerization properties
The NRs are an ancient family of transcription factors (TFs), found in metazoa and are involved in the regulation of development, metabolism, homeostasis, reproduction and cell death.
NR proteins have four domains: the N-terminal transactivation domain A/B, the DNA-binding domain (DBD) that contains two zinc-fingers and that is also involved in dimerization, the ligand-binding domain (LBD) which is responsible for dimerization and ligand binding, and a flexible hinge between the DBD and the LBD. The DBD is the most conserved domain, even among distant phyla like mammals and sponges, followed by the more variable LBD, whereas the A/B transactivation domain is not conserved even among closely related proteins.
NRs are believed to have evolved by two waves of gene duplication, one at the origin of metazoa, forming the different groups and the other at the origin of vertebrates, creating more paralogues in each group.
A classification scheme based on sequence homology of the DBD and the LBD has been developed. In humans about 50 NR genes have been characterized. These genes are organized in 18 groups, and these groups in turn are organized in 7 sub-families.
NRs bind to DNA as monomers, homodimers and heterodimers. NR homodimerizations and heterodimerizations make up a complex and highly connected network. The elucidation of the dimerization network for this family is important because the combination of different NRs in dimers increases the number of genes they regulate (combinatorial control), creates either permissive or non-

permissive dimers, combines different signalling pathways on the same promoter, and creates competition for common heterodimeric partners.
A large number of NRs are ligand-activated molecules and, because of this direct link between signalling molecules and the transcriptional response, they are prominent pharmacological targets.

Biological networks
Over the last years much attention has been focused on understanding biological systems as a whole. Such an integrative biology approach requires large volumes of data on a number of (intensive and extensive) properties of cells in reference states as well as perturbed states. Functional understanding of causal relationships is derived from the combination and interpretation of those datasets. To that end high-throughput experimental techniques have been developed and large-scale experiments have been performed (e.g., whole genome expression profiling with gene expression chips, interaction mapping for all proteins of an organism with yeast two-hybrid (Y2H) or immunoprecipitation/mass-spectrometry (IP/MS)).
Data management, integration and analysis of those data sets poses a big challenge and generic techniques for those tasks are still missing or being developed. Interpretation of the accumulated data leads to causal and functional relationship diagrams which are used to explain the phenotypic behaviour of the biological system. Some of the generic features of biological networks seem to be, e.g., the scale-free nature of protein interaction networks, where a small number of proteins are highly connected (hubs), whereas the majority are poorly connected (peripheral members). Such topologies can be found in protein-protein interaction, metabolic and genetic networks.
In our work we used the Protégé knowledge management environment to store, analyze and present data on NRs. We found that Protégé provides a natural environment for classification and dynamic (visual) exploration of a multidimensional data set. Old and new hypotheses can readily be tested by biologists without having to redevelop and redeploy new data representations. The NR dataset used for this study is available as an OWL file / Protége project.

Experimental Methods
Phylogenetic relationship of NRs:
To define the NR class hierarchy we used the classification from the Nuclear Receptors Committee (1999) which is based on phylogenetic relationships.
NR physical and genetic interactions:
Physical interaction data were obtained by two methods: large-scale semi-automated literature mining (abstract and full-text from ~59,000 Medline indexed articles on NRs), and lookup in NR reference databases.
Literature mining:
For the extraction of protein-protein interactions from literature (comprising both abstracts and full text), we used the following methodology:

1) Synonyms for every NR protein were retrieved from the AstraZeneca Gene Catalogue database and from a review paper.
2) QUOSA software was used to retrieve full-text articles and abstracts from MEDLINE that referred to NRs.
3) A keyword term (“Nuclear receptor”) was identified that would retrieve the highest number of relevant articles.
4) The following were downloaded:
   a. The most relevant 4000 full text articles of 2003 in PDF or HTML form, depending on the available form.
   b. The relevant 35353 abstracts of 11 years, from 1993 to 2003 in plain text form.
5) Full-text documents were converted from PDF and HTML into plain text.
6) Sentences containing two different NR protein names or any of their synonyms co-occurring with terms that described an interaction (*e.g.* “dimer”, “interact”) were extracted.

7) All (~3500) sentences retrieved from full text and (~2500) sentences retrieved from abstracts were read manually and those that described a physical or genetic interaction were marked.
8) Relevant sentences from the full-text and abstract subset were used for creation of the dimerization dataset.

While mining the full-text literature from 2003, we also marked any document that mentioned genetic interactions between any two NR members. Furthermore, we read the whole articles, and even followed their references, in order to verify the genetic interactions.

Gene co-expression of NRs:

Gene expression data for 49 human NR genes (using the Affymetrix HG_U133 chip data) were collected. We specifically used expression values corresponding to probe sets that mapped to an exon, or a 3'UTR in tissue samples that were classified as having normal morphology and pathology over 99 different human tissues. To determine co-expression between two NR genes we used the following criterion: a gene was "called present" if the gene transcript was called present (according to the Affymetrix MAS5 algorithm based Detection Call) in at least 50% of tissue samples for any tissue for which more than 10 experimental samples existed; if this criterion applied to two separate genes then they were assumed to be co-expressed.

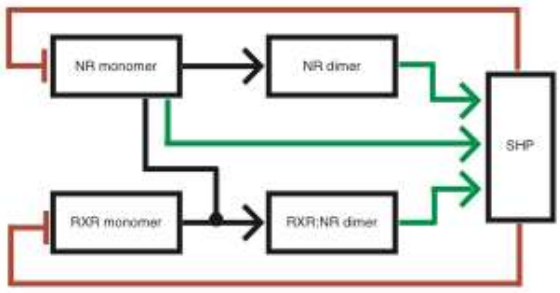
Use of Protégé:

Protégé was used as a primary data integration and analysis tool.
Protégé was employed for and particularly useful in the following task:

- data consolidation (e.g., over synonyms and alternative spellings for genes/proteins),
- data cleanup and normalization (e.g., different description of interaction and complex constructs),
- abstraction (e.g., analyzing associations between single proteins or between superclasses of proteins), and
- visual data exploration (e.g., overlay of different types of associations for hypothesis testing and elucidation of tightly coupled functional motifs).

NR Networks and Functional Relationships – Major Findings

The topology of the NR network is hub-based, but much more connected than had been thought. Similarities and differences are found in the evolution, topology and properties between two highly complex TF dimerization networks in humans, the NR and bHLH (basic helix-loop-helix) networks.
Furthermore, a large number of negative feedback loops seem to be responsible for the stability of the NR system. Finally, a very interesting crosstalk among the hubs of the bHLH and NR networks creates a large number of interdependencies between the two networks, while using a minimum number of connections.



Negative feedback loops in the NR network combining protein dimerization (black) genetic interactions (activation: green) and inhibition through protein interaction (red).

The integration of protein and genetic interactions reveals the possibility for several negative feedback loops that all share the same component: the master switch, SHP.