UNIVERSITEIT VAN AMSTERDAM

Validation of Systems Biology Models

D. Hasdemir

Summary

The paradigm shift from qualitative to quantitative analysis of biological systems brought a substantial number of modeling approaches to the stage of molecular biology research. These include but certainly are not limited to nonlinear kinetic models, static network models and models obtained by the analysis of large scale datasets such as clusters in gene expression data or principal component analysis models. However, the concept of 'model validation' is not encountered as often as the introduction of new models in the field. This leaves many of the proposed models untested and therefore, creates a gap between the number and the reliability of the models. With this thesis, we present computational approaches for model validation and provide guidelines for reliable model validation and selection with examples on real biological data.

The second and third chapters of this thesis focus on nonlinear kinetic models by which we can model the dynamics that underlie processes within a cell. They are usually formulated by using sets of ordinary differential equations (ODE) with many unknown parameters. Most of the time, there are competing hypotheses on the biochemical species, the regulatory relations that these models contain and the governing biochemical rules. This uncertainty brings the need for careful model selection and model validation.

In the **second** chapter, we present a comparative approach for model invalidation employing cross-validation which is a widely used resampling technique in statistics. Our approach is based on assessing the predictive power of a model compared to an unsupervised data analysis method, namely smooth principal components analysis. Low levels of relative predictive power indicate low informative levels of biochemical model structures that are under study. We also present results from the application of our approach on an eicosanoid production model in human and a high osmolarity glycerol pathway model in *Saccharomyces cerevisiae*.

In the **third** chapter, we extend the concept of using cross validation in the analysis of ODE based models and apply it across multiple experimental conditions. The commonly applied method for validating such models is a hold-out validation approach in which a pre-determined part of the data is used for estimating the parameters of the model and the predictions by the model on the remaining part are used to test the model structure and to select the best model structure between

competing hypotheses. However, this strategy is prone to substantial risks. The most important of them is being biased by the underlying biological facts. As an alternative, we introduce a cross validation scheme across multiple experimental conditions for more reliable model selection and validation.

The fourth and the fifth chapters focus on the validation of two conceptually different types of models, both regarding transcriptional regulation. In the first type, a regulatory transcriptional network model is used to summarize the physical association between genes and transcription factors. In the second type, clusters obtained from a cluster analysis summarize the similarity of the expression profiles of genes across various arrays.

In the **fourth** chapter, we present local and global measures to detect inconsistencies between fixed transcriptional regulatory network topologies and gene expression data. The measures we present are based on the supervised decomposition using network component analysis of gene expression data under the topological constraints defined by competing models. Competing transcriptional regulatory networks can be discriminated by using the global measure whereas unknown regulatory interactions can be identified using the local measure. Besides the simulations study through which we show the applicability of the measures, we also present the application of our method on the network model of cell cycle regulation in *Saccharomyces cerevisiae* and introduce potential points of improvement for the network model.

In the **fifth** chapter, we present an assessment of the applicability of cross validation and bootstrapping based stability measures for validation of clusters obtained from k-means clustering on large scale gene expression datasets. Additionally, we apply these approaches for the validation of clusters obtained from *Synechocystis* gene expression data and present biological results regarding the transcriptional control mechanisms regulating the day and night rhythm of the particular organism.

The final chapter of this thesis includes concluding remarks and an overview of the future perspective regarding the construction of more reliable systems biology models.

Samenvatting

De paradigmaverschuiving van kwalitatieve naar kwantitatieve analyse van biologische systemen heeft een aanzienlijk aantal benaderingen voor het modelleren van deze systemen ten tonele gevoerd in moleculair biologisch onderzoek. Deze benaderingen zijn onder andere niet-lineaire kinetische modellen, statistische netwerk modellen en modellen die verkregen zijn door de analyse van grote dataverzamelingen zoals clusters in gen-expressiedata of hoofdcomponentenanalyse modellen. Modelvalidatie heeft de introductie van nieuwe modellen echter niet kunnen bijhouden waardoor veel van de voorgestelde modellen ongetest blijven en er een kloof is ontstaan tussen het aantal modellen en de betrouwbaarheid van die modellen. In dit proefschrift worden computationele benaderingen voor modelvalidatie gepresenteerd en worden richtlijnen gegeven voor betrouwbare modelvalidatie.

In hoofdstukken 2 en 3 van dit proefschrift ligt de nadruk op niet-lineare kinetische modellen waarmee de dynamica die ten grondslag ligt aan processen in de cel gemodelleerd kan worden. Gewoonlijk worden deze processen geformuleerd in termen van gewone differentiaalvergelijkingen met een groot aantal onbekende parameters. Bovendien worden er veelal verschillende hypothesen gebruikt ten aanzien van de betrokken biochemische actoren en hun onderlinge regulatie alsook met betrekking tot de toepassing van biochemische regels.

In het **tweede** hoofdstuk wordt de in de statistiek veelgebruikte 'resampling' techniek van de kruisvalidatie gebruikt voor een vergelijkende methode om modellen te valideren. Deze methode is gebaseerd op het beoordelen van de voorspellende waarde van een model en deze te vergelijken met een data-analyse techniek zonder supervisie namelijk de gladde hoofdcomponentenanalyse. Een lage relatieve voorspellende waarde geeft aan dat er weinig informatie zit in de beoordeelde modellen. De resultaten van de toepassing van deze methode op een model voor de productie van eicosanoïde in de mens en op een model voor het hoge osmolarieteit glycerol reactiepad in *Saccharomyces cerevisiae* worden ook in dit hoofdstuk gepresenteerd.

In het **derde** hoofdstuk wordt het gebruik van kruisvalidatie in de analyse van modellen gebaseerd op gewone differentiaalvergelijkingen verder uitgebreid en wordt het toegepast voor verschillende experimentele condities. De gebruikelijke methode voor de validatie van dergelijke modellen is een 'hold-out' validatie waarbij een vooraf bepaald deel van de data wordt gebruikt voor het schatten van de modelpa-

rameters en de overgebleven data worden gebruikt voor het testen van de modelstructuur en het selecteren van de beste modelstructuur uit verschillende hypothesen. Deze methode brengt echter aanzienlijke risico's met zich mee. De belangrijkste daarvan is bias door de onderliggende biologische gegevens. Als alternatief wordt een kruisvalidatieschema geïntroduceeerd dat meerdere experimentele condities omvat waardoor modelselectie en validatie betrouwbaarder worden.

In het vierde en vijfde hoofdstuk ligt de nadruk op de validatie van twee conceptueel verschillende types van modellen die beide betrekking hebben op regulatie van transcriptie. Het eerste type model maakt gebruik van een transcriptie regulatie netwerkmodel om de fysieke associatie tussen genen en transcriptiefactoren samen te vatten. In het tweede type model worden clusters die verkregen worden via een clusteranalyse gebruikt om overeenkomsten in het expressieprofiel van genen in verschillende arrays samen te vatten.

In het **vierde** hoofdstuk worden lokale en globale maten gepresenteerd voor de detectie van inconsistenties tussen vaste transcriptie regulatie netwerktopologieën en genexpressiedata. Deze maten zijn gebaseerd op de 'supervised' decompositie van genexpressiedata met topologische randvoorwaarden die gedefinieerd zijn door coccurerende modellen, namelijk de netwerk componentenanalyse modellen. Mogelijke transcriptie regulatie netwerken kunnen worden onderscheiden door gebruik te maken van de globale maat terwijl onbekende regulatie patronen met de lokale maat geïdentificeerd kunnen worden. Naast de simulatiestudie om de toepasbaarheid van de maten te demonstreren, wordt de methode ook toegepast op een netwerkmodel voor de celcyclus regulatie in *Saccharomyces cerevisiae*. Tevens worden punten voor mogelijke verbeteringen van dit netwerkmodel aangedragen.

In het **vijfde** hoofdstuk wordt de toepasbaarheid beoordeeld van op kruisvalidatie en bootstrapping gebaseerde maten voor stabiliteit voor de validatie van clusters die zijn verkregen door k-gemiddelden clustering op grote genexpressie dataverzamelingen. Bovendien worden deze methoden gebruikt voor de validatie van clusters van *Synechocystis* genexpressie data en worden biologische resultaten met betrekking tot transcriptie controle mechanismen voor de regulatie van het dag- en nachtrithme van dit organisme gepresenteerd.

Het laatste hoofdstuk van dit proefschrift omvat afsluitende opmerkingen en een toekomstperspectief ten aanzien van de constructie van meer betrouwbare modellen in de systeembiologie.