G.A. nº 270086

Collaborative Project of the 7th Framework Programme

WP5: Multi-scale horizontal integration

Deliverable 5.1: Horizontal integration: Strategies for connecting mechanistic and probabilistic modelling

[v.6] [31.08.2011]

www.Synergy-COPD.eu

Document Information

Project Number	270086	**Acronym**	Synergy-COPD
Full title	Modelling and simulation environment for systems medicine (Chronic obstructive pulmonary disease -COPD- as a use case)		
Project URL	http://www.Synergy-COPD.eu		
EU Project officer	Marta Lorens		

Deliverable	**Number**	D5.1	**Title**	Horizontal integration: Strategies for connecting mechanistic and probabilistic modelling
Work package	**Number**	WP5	**Title**	Multi-scale horizontal integration

Date of delivery	**Contractual**	PM6	**Actual**	PM7
Nature	Prototype ☐ Report ☑ Dissemination ☐ Other ☐			
Dissemination Level	Public ☑ Consortium ☐ Restricted ☐			

Document responsible	David Gomez-Cabrero		**Email**	david.gomezcabrero@ki.se
	Partner	KI	**Phone**	+46 (8) 517 708 69

Authors	**Name**	**Partner**
Main author(s)	David Gomez-Cabrero	**KI**
Co-author(s)	Isaac Cano	**IDIBAPS**
	Luigi	**BDIGITAL**
	Dieter Maier	**BIOMAX**
	Peter Somogyi	**KI**
	Michel Mickael	**KI**

Abstract (for dissemination)	The general aim of Synergy-COPD is to develop “a simulation environment and a decision-support system aiming at enabling deployment of systems medicine”. To this end the Consortium will make use of existing knowledge in the form of already available mechanistic models. In addition, we aim to integrate in the simulation environment novel information representing the "relationships between entities" that are derived by the application of network inference approaches.
Key words	Probabilistic models, mechanistic models, network inference, datasets.

Version Log			
Issue Date	**Version**	**Author**	**Change**
10.08.2011	V1	David Gomez-Cabrero	draft
22.08.2011	V2	Luigi Ceccaroni	revision
25.08.2011	V3	David Gomez-Cabrero	Structure modification
26.08.2011	V4	Consortium	Version review by consortium members
31.08.2011	V5	Francesco Falciani	Review of references
31.08.2011	V6	David Gomez-Cabrero	Update of references. Last version

Index

1 introduction

1.1 Inference Approach. Network Inference methods of relevance to Synergy-COPD

The general aim of Synergy-COPD is to develop “a simulation environment and a decision-support system aiming at enabling deployment of systems medicine”. To this end the Consortium will make use of existing knowledge in the form of already available mechanistic models. In addition, we aim to integrate in the simulation environment novel information representing the "relationships between entities" that are derived by the application of network inference approaches (e.g. information theoretical approaches such as ARACNE, Basso et al. (2005)) to a number of available datasets (see Section 1.1 below for a detailed description of such data sets).

Deliverable 5.1 describes the approaches we plan to follow in order to achieve this overall objective. Section 1, includes the introduction that describes the overall goal, and the concepts and state of art behind the mechanistic models (MM) and the Inference Approaches (InfA) considered. Far from a exhaustive description we focus the presentation on the theory and methods relevant for the project.

Section 2 describes MM and InfA within the Consortium. First subsection explains the development of network inference methods (nIM); those nIM show novel important candidate pathways including non-previously modeled biological processes of relevance (for example tissue remodeling). Second subsection describes the mechanistic models under consideration.

Section 3 describes the approach that aims to integrate the relevant information that can be extracted from MM and InfA approaches within a unifying framework. We provide details about the development of several candidate methods. Synergy-COPD Consortium is already running several analysis and integrative approaches, therefore sections 2 and 3 of this document include the research proposed and, at some points, the research done up to August 2011.

Finally section 4 provides some conclusions and open questions.

1.2 Mechanistic approach. The theoretical background in use.

Mechanistic approaches make uses of mechanistic models to extract knowledge and/or to generate predictions. In this section modelling properties, purposes and types are covered. Then steps of building a model will be introduced with an emphasis on the mechanistic model and its ODE structure.

It's important to introduce the concept of modelling before focusing on the mechanistic modelling approach. In this sense, one could assume that two different point of views exist the first is the abstract world, which is built out of ideas, whereas the second world is based on observation and data. These two worlds are linked through an iterative cycle, which is based on improving the ideas by doing experiments that produce data, then improved ideas need improved data, and the cycle continues. In this sense, Modeling could be defined as the art of improving ideas based on data (Szallasi Z et al (2006)). This is practically done through defining set of rules that maps a set of inputs to a set of outputs. In this case A model could be interpreted as a representation of the construction and working of some system of interest (a biological one in our case) (Szallasi Z et al (2006)) .

Normally a model is similar to and simpler than the system it represents. It usually consists of variables, parameters and functional forms (Ellner SP et al (2006)) . One purpose of a model is to enable the analyst to predict the effect of changes to the system. Also a good model should reflect the known experimental data. Analysing the model could be used to understand which parts of the system contribute most to certain desired properties of interest. This would lead to the formation of new hypothesis and the ability to analyze the effects of manipulating experimental conditions in the model without having to perform complex and costly experiments (or to restrict the number that is performed) (Bender EA,(1978)), (Cross M (1985),(Fowkes ND et al,1994))

On one hand, a model should be a close approximation to the real system and incorporate most of its important features. On the other hand, it should not be so complex that it is

impossible to understand and experiment with it. A good model is a compromise between realism and simplicity (Maria A (1997)).

Next paragraphs review the process of building, analysing and validation a mechanistic model. By mechanistic model we understand the model that describes a system by its constituent parts and mechanisms.

Steps of building a model'

It is defined as a series of steps taken to convert a crude idea into a model. First a conceptual model is formed then a quantitative model is formalized. A conceptual model represents ideas about how the system works. A quantitative model involves equations that are developed for intrinsic processes and then combined to form a system of equation in case of complex systems.

To build a quantitative model, the first step is the choice of the inputs and outputs. The logical second step is the choice of the range of the inputs and then defining the precision or the accuracy followed by the step of defining measurements types and numbers (Selinger DW (2003)). This idea is shown in Figure 1.

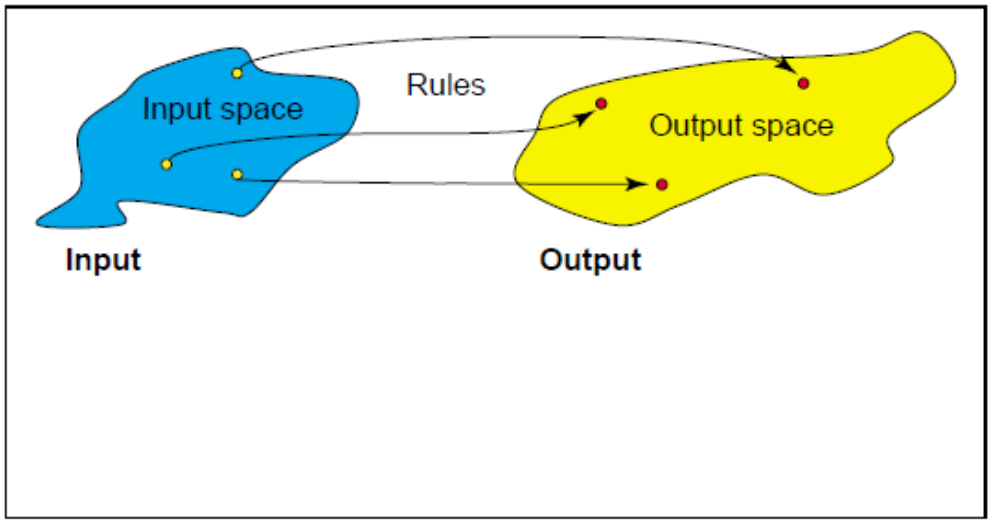


Figure 1: A model map certain set of inputs into a certain set of outputs, using a set of rules(Selinger DW (2003)).

Then comes the step of formulating the assumptions; this normally reflects the beliefs about how the system operates. Future analysis of the system treats these assumptions as being true, but the results of such an analysis are only as valid as the assumptions. If the assumptions are sufficiently precise, they may lead directly to the mathematical equations.

A wide variety of natural phenomena such as projectile motion, the flow of electric current, and the progression of chemical reactions are described by equations that relate changing variables. As the derivative of a function provides the rate at which that function is changing with respect to its independent variable, the equations describing these phenomena often involve one or more derivatives, this is usually referred to as differential equations. Types of equations: there are several types of ordinary differential equations which could be used to represent a certain system depending on the complexity, state and the nature of the modelled system.

- First Order Differential Equations and it includes separable variables, homogenous, exact, integrating factor, linear and Bernoulli equation (Larson RE et al,(1994)).

Summary of First-Order Differential Equations

Method	*Form of Equation*
1. Separable variables:	$M(x)\,dx + N(y)\,dy = 0$
2. Homogeneous:	$M(x, y)\,dx + N(x, y)\,dy = 0$, where M and N are nth-degree homogeneous
3. Exact:	$M(x, y)\,dx + N(x, y)\,dy = 0$, where $\partial M/\partial y = \partial N/\partial x$
4. Integrating factor:	$u(x, y)M(x, y)\,dx + u(x, y)N(x, y)\,dy = 0$ is exact
5. Linear:	$y' + P(x)y = Q(x)$
6. Bernoulli equation:	$y' + P(x)y = Q(x)y^n$

- Second Order Differential Equations and it includes nonlinear equations, linear equations, homogenous linear, homogenous with constant coefficients, non-homogenous and series solutions (Stewart J (2008)).
  The second order differential equation takes the form of:

$$\frac{d^2y}{dx^2} = f(x, y, y')$$

- Systems of Differential Equations and they involve more than one unknown function.

If a model is defined by a set of differential equations, sometimes it is possible to find an analytical solution; however this is most of the times not the case.
One way to get around this is the use of a numerical approach that allows us to solve the system under certain conditions for a given certain initial conditions and parameters. However, in this case the solution is approximated and error estimation is needed; in many cases it is done through using an iterative approach. A well-known numerical method is Rung-Kuta, which could be used to solve a host of different mathematical problem (Butcher JC(2003)).

Incomplete information in a model

In most cases, the parameters of a system and/or the structure of the system is not completely known. In those cases experimental data is generated to give values to unknown parameters: optimization techniques together with ensembling approaches are widely used to compute unknown parameters (Avriel M (2003), Frederick BJ et al (2006), Frederick BJ et al (2000), Boyd S et al,(2004), Nocedal J(2006)).

Behaviours in a model

It is important to realize that the behaviour of a model can be described in two ways. Qualitative description provides an answer to questions about “how”, whereas quantitative description answers questions about “how much”. The study of both descriptions is relevant.

Sensitivity analyses

Sensitivity analysis is a broad concept that has been used across various disciplines. It can be used to identify a risk assessment of a certain procedure (model). Here risk assessment is defined in a general sense to mean the probability of a certain output of an equation. For

example in the case of a model with bi-stable conditions (such as the one presented in (Selivanov et al (2009),Selivanov et al (2011)).

The sensitivity analysis could solve different questions such as: what parameters help to achieve a certain state?, and in what way reaching a certain state variable is related to a certain parameter (concentration)?

From a broad point of view, risk assessment can be used to determine which parameters are essential to control, to reduce, or to eliminate from the original model. Therefore, risk assessment can help in developing more effective COPD control plans.

The objective of the sensitivity analysis is to answer the following question (Frey HC et al,(2002),Kaplan S et al(1981))

- What happened:

- How did it happen

- How likely is that to happen again

Risk assessment of a certain process or a model based could be briefed in the following for steps:

- Parameter Identification: for instance the identification of the parameters that most control different processes.
- Parameter Characterization. This could be defined as the evaluation of the nature of the effects caused by changing these identified parameters.
- Exposure Assessment. It is evaluation of the likely existence of such biological parameters.
- Risk Characterization. This involves estimation, of the probability of occurrence and severity of a certain output.

Sensitivity analysis methods can be classified as:

(1) Mathematical sensitivity analysis methods: They assess sensitivity of a model output to the range of variation of an input. These methods typically involve calculating the

output for a few values of an input that represent the possible range of the input These methods also can be used for verification and validation.

(2) Statistical methods: They are done through running simulations in which inputs are assigned probability distributions and assessing the effect of variance in inputs on the output distribution where one or more inputs are varied at a time. Statistical methods identify the effect of interactions among multiple inputs. The range and relative likelihood of inputs can be propagated using a number of methods such as Monte Carlo simulation, Latin hypercube sampling, variance, response surface methods, Fourier amplitude sensitivity test, and mutual information index.

(3) Graphical methods: It is concerned with representation of sensitivity in the form of graphs, charts, or surfaces. Generally, graphical methods are used to give visual indication of how an output is affected by variation in inputs. Graphical methods can be used as a screening method before further analysis of a model or to represent complex.

To the interest reader we refer to **Annex 1. Sensitivity Analysis methods.**

Validation

Once the model has been developed then the issue of validity arises. This is usually done through simulating the model under known input conditions and comparing model output with known system output. One of the advantages of using mathematical modeling is enhancing scientific understanding, as this type of modeling usually embodies a hypothesis about the study system, and it also allows comparing this hypothesis with the available data. Also mathematical models are useful experimental tools for building and testing theories, assessing quantitative conjectures, answering specific questions, determining sensitivities to changes in parameter values and estimating key parameters from data

1.3 Integration within Synergy : what and how?

It is important to define the boundaries and goals of the integration. Where the network

inference methods aim to discover the hidden structure by the use of data, the mechanistic models aim to show the behavior of selected variables by the use the causality knowledge of a system (where some of those variables cannot be observed by experimental analysis). We define our goals as the use of those two approaches within the same knowledge discovery approach.

For instance, the mechanistic models, which will be part of the simulation environment, represent reasonably well-understood biological processes such as oxygen diffusion in the lungs and muscle bioenergetics, including the production of reactive oxygen species. Other important components of muscle physiology (for example, the oxygen-dependent transcriptional control of energy-related enzymes and the activation of tissue remodeling pathways) are less understood at a mechanistic level but are nevertheless important to model muscle wasting. It is an objective of Synergy to include these additional biological processes in the simulation environment. This objective will be achieved by 'learning' the overall structure of the unknown biological networks using network inference methodologies. In this example the Integration aims to use available data to extend the mechanistic models. This "extension" of the causality can lead to new discoveries that no approach would be able to decipher separately.

To summarize we consider (1) the use of inference network analysis to extract the causality relations among variables in data sets and (2) the use the mechanistic models to use the know causality to infer variabless´ values (that are usually non-observable) and to generate predictions. The evident use is that under those considerations the output of (1) can be considered and the input of (2) and viceversa.

However our aim is not to describe the use and integration of different inference network approaches.

2 InfA and MM approaches in Synergy.

Being the major goal of this deliverable to integrate the information provided by two different approaches, PA and MA, it is then necessary to describe them to certain degree of detail within Synergy-COPD environment. First sub-section focuses in PA and it describes the use

of the network-inferring and reverse engineering strategy that has been developed and implemented by BHI; BHI's research has focus in BioBridge database, however other resources are to be considered in the near future, therefore we describe them to certain detail, with an emphasis in the analysis to be performed. Second sub-section focuses in the MA approach and it describes the models to be considered and the information we aim to extract from their use. Deliverable 4.1 (D4.1) described in more details each one of the models; in D4.1 three mechanistic models are considered:

- M6. Oxygen transport and utilization (M1 + M2)
- M3. Spatial heterogeneities of lung ventilation and perfusion
- M7. Bioenergetics, and mitochondrial respiration and reactive-oxygen-species generation (M4 + M5)

Which are an evolution and integration of five previously considered models:

- M1. Central and peripheral oxygen transport and utilization
- M2. Pulmonary gas exchange
- M3. Spatial heterogeneities of lung ventilation and perfusion
- M4. Skeletal muscle bioenergetics
- M5. Mitochondrial Reactive Oxygen Species (ROS) generation"

Note that Section 2 will be updated as soon as new analyses are done and as soon as results are ready for clear presentation.

2.1 The application of network inference approaches in Synergy-COPD

The mechanistic models that so far are of the Synergy-COPD simulation environment link oxygen diffusion from the lungs to energy metabolism in the muscle and focus on the production of Reactive Oxygen Species (ROS) as a clinically relevant endpoint (this will be

further discussed in section 2.2). The role of ROS in muscle wasting is well supported by the literature Barreiro et al. (2009) and clearly represents an important mechanism. However, there are a number of additional factors that need to be taken into consideration for the simulation platform to reflect the complexity of a muscle.

The bioenergetics models to be integrated in Synergy-COPD are based on the assumption that the rate of many reactions is invariant. This assumption is not completely correct. We know for example that the expression of several enzymes involved in glycolysis and gluconeogenesis is likely to be influenced by oxygen availability (represented by VO2max, Turan et al. (2010)) and we recently demonstrated that they can be affected by cytokine and growth factors produced in the muscle in response to training, Turan et al. (2010). In addition, muscle wasting is likely to be the result of the failure to activate tissue-remodeling pathways in response to physical training. These are essential components of muscle homeostasis and are themselves influenced by oxygen availability and cytokine and growth factor signals in the muscle.

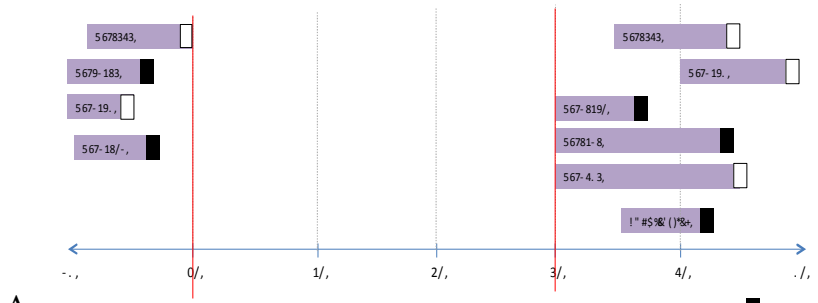
Although the exact mechanisms behind oxygen and growth factor mediated regulation of bioenergetics and tissue remodeling is not fully understood, it is possible to infer high-level models representing these relationships from observational data. Here we describe approach, datasets and network inference methodologies we plan to use to achieve this objective. Finally we will discuss the strategies for integration of the inferred models within the Synergy-COPD simulation environment.

2.1.1 Initial data set considered

Integration and Meta-analysis of Expression profiling datasets, representing human healthy and diseased muscles

The availability of large datasets representing the transcriptional state of healthy and diseased muscles is key to our efforts in network inference. For this reason we decided to complement the existing expression profiling data from the Framework VI BIOBRIDGE project with additional datasets available in the public domain. Therefore, seven additional datasets were identified to represent two very distinct age groups, one of which linked to the age of first

diagnosis of COPD. Figure 2 represents a schema of the age distribution of the individual datasets, the microarray platforms used in the original studies and the number of samples in each group. The dataset have been already integrated using the statistical methodology Combat, Johnson et al. (2007) Individual and the integrated datasets are already stored in the Biomax knowledge management system.



!"#$%&	'("&*+("&	!*, --"%&	'("&*'0'(1#2&
5678343,	9/%0/:,32%42,	-1:;-3,	;<=>?:,")@
5679-183,	-.A0%924,	4,	;<=>?,
567-19.,	-.%92:,4/%/,	-/:;-9,	;<=>?:,")@
567-18/-,	-8%91,	91,	;<=>?,
567-819/,	1.%81,	-9,	")@
56781-8,	3/%41,	--,	")@
567-4.3,	23%43,	-9,	")@
!"#$%&'()*&+,	3/%82,	91,	")@

B

Figure 2. Integration of existing datasets representing muscle biopsies from normal individuals in two different age groups

Mouse models of hypoxia and physical training

The analysis of the human muscle biopsies integrated datasets (see above) will give us the possibility to infer gene-to-gene relationships linking for example VO2max and cytokine signaling to the expression of enzymes involved in bioenergetics and to the activity of pathways involved in muscle homeostasis. However, in order to develop computational

models representing the dynamics of response to physical training in hypoxia and normoxia an additional dataset will be acquired from ongoing efforts in Dr. Falciani's laboratory. In order to be able to work in fully controlled conditions a mouse model of chronic hypoxia will be used. Mice in normal environment (normoxia) and mice in hypoxic chambers will be subject to the same training regime obtained using surgically implanted and remote controlled electrical stimulators (see Figure 3 for details of the experimental design).

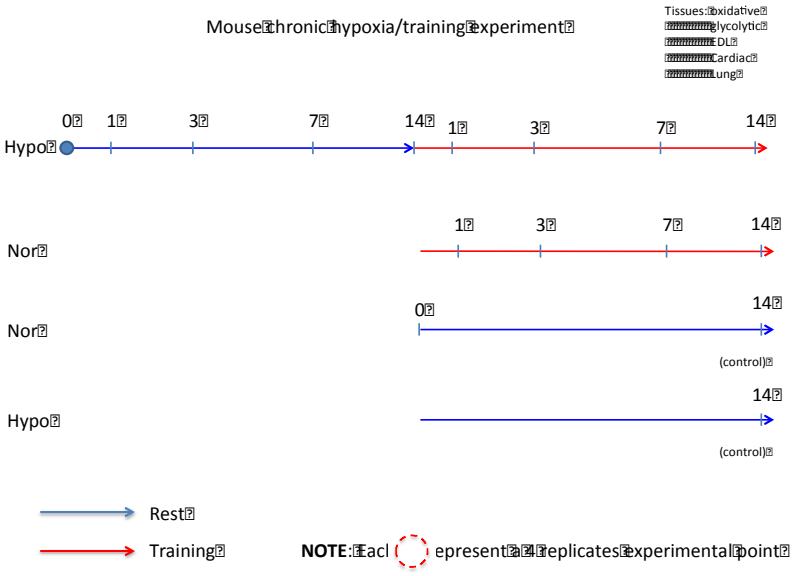


Figure 3. Mouse training experiments: Hypoxia effects.

2.1.2 Network inference strategy: A graphical model representing young and old muscles.

As introduced above, the main goal of network inference will be to develop a graphical model linking important physiological parameters (for example, VO2max and indicators of systemic and local inflammation) to the expression of enzymes involved in bioenergetics and to the

activity of pathways involved in muscle homeostasis.

At a methodological level we will first apply information theoretical approaches based on mutual information (ARACNE, Margolin et al. (2006)) on the compendium of gene expression profiling data described in section 2.1. In addition to ARACNE, we will also use linear and non-linear measures of correlations and statistical model fitting for defining the exact shape of the relationship and its sign (positive or negative).

Graphical models representing young and old muscles will then be developed by thresholding the interaction matrixes using a *p-value* threshold estimated using a bootstrap approach and visualized as a graph in the software application Cytoscape. Individual network modules will be identified using MCODE (Bader and Howe (2003) and tested for functional enrichment using the web-based application DAVID, Huang et al. (2009).

Physiological measurements such as VO2max are not available for all datasets; therefore we will not be able to directly link these measurements to the networks. However, network nodes that are correlated to VO2max can be identified using the BIOBRIDGE 8-weeks training dataset (Turan et al. (2010)) and then mapped on the previously inferred network.

2.1.3 Inferring the relationships between VO2max and the molecular state of healthy and diseased muscles

As mentioned above, one of the key objectives of the analysis will be to identify relationships between oxygen availability, expressed as VO2max, and mRNA expression of enzymes involved in **energy metabolism**. Although this cannot replace protein measurements or measurements of enzyme activity it will give us the possibility to estimate which components of the bioenergetics model can be affected by different levels of VO2max. Since VO2Max varies within the population of COPD patients and changes in response to intervention the inferred relationships will be relevant for both steady state and dynamic models. Of particular interest will be the connection between VO2max and network modules representing the activity of **pathways involved in muscle homeostasis**. Relevant examples of these are production of structural components of muscle fibers and angiogenesis. As we have discovered that genes involved in these pathways are highly correlated with each other we

will develop **indexes of pathway activity** rather then relying on individual gene measurements [see Antczak et al. (2010) for a description of the procedure].

Inferring the relationships between molecular indicators of response to training and the molecular state of healthy and diseased muscles.

We previously defined that the expression of a subset of cytokine and growth factor receptors (IL1R, VEGFA, CSF1 and ENO1) involved in muscle response to stimulation correlates to the expression of a number of key components, involved in muscle remodeling. Here we plan to apply the same methodologies described previously to identify the statistical relationship between these markers and bioenergetics and tissue remodeling functions. In modeling response to training in COPD this is likely to be particularly important since only a fraction of these patients is responsive. The availability of markers will therefore allow to customize each simulation to specific patients.

Developing dynamical models from the inferred networks

The sub-networks identified by the strategy outlined above will represent static relationship rather than the temporal evolution of molecular changes following intervention. In order to develop dynamical models parameters will need to be estimated from relevant datasets. Here we propose to use the mouse training datasets in conjunction with a computational framework for inferring dynamical networks from observational data, Gupta et al. (2011). The learning process will be driven by the knowledge of the relationships but the exact value of the inferred parameters will depend on the nature of the relationships inferred by the procedures outlined before.

Although this will not be the final form of the model to incorporate within the simulation environment, it will constitute a good proof of concept.

2.1.4 Extending the approach to other data-sets

In addition to BioBridge, Synergy-COPD considers two other databases: PAC-COPD and

Eclipse-GSK. At the end of the project, validation will be implemented using Eclipse-GSK, because it is similar in nature to PAC-COPD.

PAC-COPD[1] "is a cross-sectional and cohort study of 342 patients with COPD from 9 tertiary hospitals in 3 autonomous communities. The minimum follow-up period is 5 years. The main variables of interest are respiratory symptoms, smoking, alcohol use, physical activity, use of health care services, medical care, treatment received, activities of daily living, comorbid conditions, sleepiness, anxiety and depression, quality of life, forced spirometry and bronchodilator tests, lung volume and inspiratory capacity measured by body plethysmography, carbon monoxide diffusing capacity, baseline arterial blood gas values, respiratory and peripheral muscle function, electrocardiogram, body weight and composition measured by bioelectric impedance, chest radiograph, skin prick test, capacity for exercise measured in the 6-minute walk test and cardiopulmonary exercise test, induced sputum (for quantitative microbiological culture and determination of inflammatory markers), nighttime pulse oximetry, chest computed tomography scan, and echocardiography. Levels of markers of inflammation and oxidative stress are measured in serum and plasma; these samples are also used for genetic analysis and will be stored for other possible measurements that might be required in the future."

Relevant published analysis using PAC-COPD include the following results:

- "In patients with COPD recruited at their first hospitalization, three different COPD subtypes were identified and prospectively validated: 'severe respiratory COPD', 'moderate respiratory COPD', and 'systemic COPD'." (Garcia-Aymerich et al. (2009))

- "Between a third and a quarter of patients with clinically stable COPD present abnormal levels of circulating antinuclear and anti-tissue antibodies, the latter being related to lung function impairment. These observations provide further support to the hypothesis that the pathogenesis of COPD involves an autoimmune component." (Nunez et al. (2011))

- "A significant association between anxiety, depression, or both conditions and impaired HRQoL. Clinically relevant factors affecting the magnitude of this association include work status, COPD severity, and the presence of comorbidities." (Balcells et al (2010))

- "Moderate-to-severe Spanish COPD patients report an adequate intake of the main food groups and macro- and micro-nutrients according to local recommendations, excepting vitamin D". (de Batllet et al. (2010))

- "Patients admitted after presenting with their first COPD exacerbation have a wide range of severity, with a large proportion of patients in the less advanced COPD stages." (Balcells et al. (2009))

Statistical analysis has been performed on the PAC-COPD database (i.e. ref. 1 reconsiders the definition and classification of COPD, which is still under investigation). However, our aim is to use the database to study questions related to causality (the mechanistic understanding of the disease development) and prediction (to use the data to predict prognosis and asses the situation of a patient); both ideas are developed in Section 3.

To allow for such analysis and in order to succeed in the integrative approach, several steps need to be taken (some of them are near completion):

Step 1: Annotation of the models. All models have been annotated (see deliverables 4.1 and 3.1).
Step 2: Selecting sections of the PAC-COPD database that can be related to the data from BioBridge and/or to the models M6 and M7. Annotate those sections following similar procedures to the ones used to annotate models and BioBridge.
Step 3: Use annotation above to map elements between PAC-COPD and Biobrdige-M6-M7.

Once this is done, the selected parts from PAC-COPD will be ready for analysis by statistical methods and by the procedures mentioned in Section 3 of this deliverable. **The very first basic idea is to find relations (such as correlation) between elements within each data set and among data sets** (as the "subnets"/contexts mentioned in D3.1 section 3.1.1).

2.2 Mechanistic model

In this section we describe the different models under consideration within Synergy-COPD and the type of analysis that we aim to generate.

We are considering M6 and M7, to be used in the integrative approach. However we consider different approaches to extract information from those models. First those models can be considered separately; secondly we can use the model that integrates both M6 and M7 within a single framework.

We have omitted M8 in the integrative approach as it is under development. Once this model becomes available for study this sub-section will be updated to include it if necessary.

2.2.1 The analysis of the models independently

As mentioned before there are two models, denoted M6 and M7. We describe first each model separately; next we detail the mathematical analysis we expect to do in each one of them to highlight the characteristics of interest. Finally we provide a review of how to use databases in the analysis of the models.

M6 model.

M6 consists of 5 main equations and it covers the relationships of oxygen pressure and volume from the lung to the mitochondria. There are two inputs, VO2max, which is the amount of oxygen consumed and p50mitochondria, which is the partial pressure of the oxygen. The output is in the form of different pressures and oxygen gas characteristics, specifically given by:

- oxygen delivery
- oxygen extraction

- mean oxygen and co2 pressures in the capillary
- difference between oxygen (and co2) pressure in the capillary and the tissue
- amount of oxygen and carbon dioxide

Further details of the model are provided in Section 3.1 of D4.1.

M7 model

M7 model integrates M4 and M5 models that were covered in 2 papers. The first paper covers complex 3 of the electron transport (see Selivanov et al. (2009)). The main aim of the paper is to model the stability condition, which arises from increasing the concentration of succinate. The second paper added complexes 1,2, and 4, kerb cycle and hydrogen transfer were added (see Selivanov et al. (2011)

Further details are provided in Section 3.3 of D4.1.

Mathematical analysis of a model.

Given a model M6 (similarly for M7), we define Parameters of Interest (PI) the subset of parameters that are considered of interest. Similarly we define State Variables of Interest (SVI). For a given instantiation of a model (a defined values for parameters and inputs) we consider the analysis of the following relations (R):

- R-PI-SVI: Relations between each PI and each SVI.
- R-SVI-SVI: Relations between pairs of SVI.

In all cases we define those relations by:

- The shape of the relation: linear, monotonic, quadratic,…
- Sensitivity analysis: by measuring how the modification of one PI or SVI affects another SVI. Different grades of strength can be defined for these relations. (See

Section 1.2 and Annex 1).

The analysis is already in process. For instance, in the case of M6, we have done the following (and further sensitivity analysis is underway):

- Considered the effect of using large quantities of oxygen partial pressure at the mirochondria at saturation of 50% (pm50) on the output, as it was previously considered to be null. Through this certain relationships between the two inputs and the outputs were identified; for instance: (1) the relationship of oxygen delivery and oxygen extraction is inverted and (2) increasing pm50 will increase oxygen delivery and hence decrease extraction.

- One potential step further of this analysis was the study of the model by changing the two inputs at the same time. This was done by forming two loops the first one for vo2max and the second for pm50. Certain results were observed, for instance, it was clear that increasing VO2max increases extraction, while it decreasesoxygen delivery.

This is also done in accordance with the deliverable 4.1, where analyzing the parameters for the above model can give insights to the outputs that might affect the production of the reactive oxygen species. One of the potential candidates in these parameters is the po2 for the mitochondria.

The analysis of M7 will contain an extra step as to define part of the parameter values a parameter optimization problem is solved by using Simulated Annealing to minimize the deviation from measured dynamics of NAD. We aim to analyze sets of solutions by defining a quality threshold.

Extending the analysis of a model by data from databases.

Parameters and/or state variables can be identified in some of the databases. Therefore the values can be used to consider the physiological ranges for healthy and COPD individuals and analyze the properties of the models for each case. Let us consider several scenarios:

Scenario 1: There is a set of PI which are measured in a given database. For healthy and COPD patients the distribution of parameter values are the same. In this case we analyze R for all individuals.

Scenario 2: There is a set of PI which are measured in a given database. For healthy and COPD patients the distributions of parameter values are different. In this case we analyze the relations R-*PI-SVI for the different populations and analyze its differences.*

Scenario 3: A mixture of Scenario 1 and 2. We propose the identification of those parameters whose distributions are different for healthy and COPD patients and to apply Scenario 2´s approach.

2.2.2 The analysis of the integrated model

In any case we propose similar approach as the presented in 2.2.1 but using the integrated model M6+M7. The architecture considered to run the integrated model will be defined in D4.2; D4.1 mentions the manually integrated model that we are using by now. In both cases, the following comments apply.

The model of mitochondrial respiration and ROS production considers in detail the intracellular mechanisms defining the energetic status of the cell and the levels of oxidative stress. Until now the application of this tool was restricted by the basic studies of intracellular processes. The application of a model is defined by the experimental data, which it can simulate. The model of mitochondrial respiration simulated the biochemical reactions of TCA cycle, the changes of mitochondrial transmembrane potential as a consequence of changes in electron transport from NADH and succinate to oxygen and proton translocation. This model described also the details of ROS production as a side product of electron transport, which has an important role as a signal or damaging factor. Respectively, it can be used for the deep analysis of the experimental measurements of oxygen consumption by mitochondria, changes in the concentration of metabolites of TCA cycle, transmembrane potential, ROS production. The analysis of such processes gives an understanding of the

state of cellular energetics under the normal conditions or under stress or pathology. Thus, if such a model would have been applied to the study of cell operation in the state of disease, it could provide important information about the basic mechanisms of pathological changes in bioenergetics under the disease conditions. In order to extend the area of application of the model to medicine, the model must be modified, it must have access not only to the biochemical experimental data, but also to the clinical data.

One of the most important factors for cellular bioenergetics, and, in the same time easy measurable in clinical conditions, is oxygen consumption. Oxygen consumption can be measured by controlling gas contents in expiring air, or by measuring of oxygen saturation of arterial and venous blood. If oxygen delivery from atmospheric air to mitochondria is described correctly, the model analysis can predict the state of intracellular energetics, which is important for understanding the biochemical basics of a number of pathologies and the development of effective therapy. The model of oxygen delivery from lungs to tissues developed by Peter Wagner can provide a link between the analysis of cellular bioenergetics and clinical data of oxygen consumption. Therefore the link of mitochondrial model with the model of oxygen delivery is necessary to analyze clinical data of oxygen consumption and understand the biochemical consequences of impaired oxygen delivery in specific cases of diseases such as COPD.

One of the ways of integration of models of mitochondrial respiration and oxygen delivery is described in Section 4.1 of deliverable 4.1. Briefly, the point of connection of the two models is oxygen concentration in cells (a variable that is included in both models). Mitochondrial model describes its consumption; blood circulation model describes its delivery. The dynamic equations of mitochondrial model give the value of consumption in each subsequent moment starting from some initial value. The circulation model has analytical solution and can be presented by a system of five algebraic equations. If these equations are solved at each time step of numerical solution for mitochondrial respiration, the joint solution of both models provide oxygen concentration in the cells and he state of cell bioenergetics as a function of the capacity of an organism to deliver oxygen.

After performing such a integration the model can be used for the analysis of patient respiration and predicting the state of intracellular energetics.

The analysis of the integrated models follows the same process as the presented for each individual model. However we will compare the results of the analysis of the models before and after integrations.

3 Integration of mechanistic and probabilistic models: "A Quest for causality and prediction".

Most of the methods previously described focuses in either "large data sets and the inference of networks"; or in the analysis of mechanistic models that reflect the relation between different parameters and state variables and/or their predictive power. However the major goal of this deliverable is to provide a unifying approach for both frameworks. Our aim is to generate integrative approaches that enhances our understanding of COPD´s progression causality (mechanistic driven approach) and/or allows us to develop better prediction methods on patient situation and prognosis.

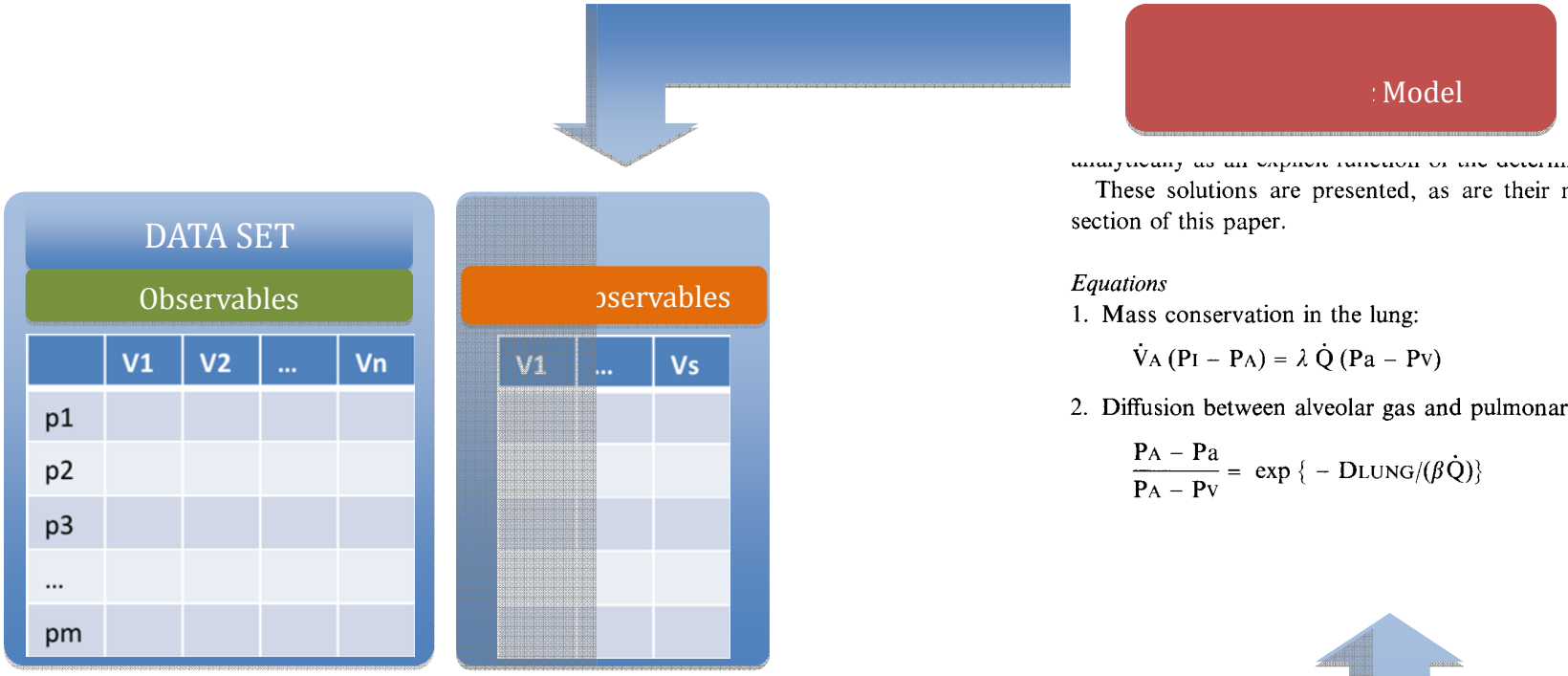
The first step in this integration is to describe the links from both approaches, and how each one is related to the other:

- **InfA to MM (Inference Approach to Mechanistic Approach)**: the final goal of an inference approach is to decipher the inner structure within a data set and the relations between their variables. This structure can be described by a graph (such as DAG graphs described in Section 1.1) that states relations between variables. However *one* major final aim of those inference methods is to decipher the "causality relations" between elements of the data set, which is to decipher the mechanistic model behind the data set. These new causality relations can be integrated or added to existing models.

- **MM to InfA (Mechanistic Approach to Inference Approach)**: a mechanistic model describes the causal relations among the different variables in the model. Some of those variables are usually observables that can be consulted also in data sets, however in many cases that is not the case. One of the aims of a mechanistic model

is to describe the values and or behavior or certain variables that are hardly observable and are not appearing in data collections. In this sense, MM can be used to extend large data sets by including “new columns” that determine the values of those non-observables. The extended large data sets can be then further studied by InfA methods. Section 3.2 describes this proposal in detail.

Figure 4 describes both relations. From this figure it is obvious to realize the existence of a loop. We describe in more detail the specific options we consider on how to use these relations, but it is evident that the synergy of this approach lies in the discovery loop highlighted. The proposed loops are described with specifications in the following two sub-sections. In all cases the mapping between data-sets, InfA outputs and mechanistic models are essential; or this reason we recommend a careful reading of deliverable D3.1 where the mapping efforts are being described.



These solutions are presented, as are their nu
section of this paper.

Equations

1. Mass conservation in the lung:

$$\dot{V}_A (P_I - P_A) = \lambda \dot{Q} (P_a - P_v)$$

2. Diffusion between alveolar gas and pulmonary

$$\frac{P_A - P_a}{P_A - P_v} = \exp \{ - D_{LUNG}/(\beta \dot{Q}) \}$$

3.1 InfA to MM

The generic aim is to extend our understanding of causality, which is to integrate the "causal knowledge" by integrating the InfA "causal output" and the MM model under consideration. In this sense, several scenarios can be considered dependent of InfA´s outputs.

Scenario 1:

InfA output: a static network that highlights relations between variables. Those relations are not necessarily causal.

Proposed action:

(1) To study those elements that are close (by network distance definition, that will depend on the network definition) to "many or all" of the state variables/parameters of MM model.

(2) Select candidates to be added to MM model.

(3) Consider the benefits of this "adding". Expert knowledge (clinicians) is required here.

(4) If the benefits exceed the costs we add the elements to the model. In this case the trade-off is hard to define as the specific causal relation among variables is not given by the InfA approach, but only that there is a relation. Only if the causality can be well defined based on expert knowledge and it gives relevant input into COPD analysis, we would consider to extend actual models.

Scenario 2:

InfA output: a static dynamic network that highlights "causal" relations between variables.

Proposed action:

(1) To study those elements that are close (by network distance definition, that will depend on the network definition) to "many or all" of the state variables/parameters of MM model.

(2) Select candidates to be added to the MM model, and the sub-networks related to them.

(3) Consider the benefits of this "adding". Expert knowledge (clinicians) is required here.

(4) If the benefits exceed the costs we add the elements to the model. In this case we have a causal relation clearly defined. However causality is explained in different ways by different methods. We consider two sub-scenarios:

Case (*a*): InfA´s ouput can, in some cases, be explained as a probabilistic causal relation (such as bayesian networks or statistical models). Then it is possible (1) to consider transforming this knowledge into a ODE system and then to do the integration or (2) to transform the ODE system into a network and then to do the integration. In both cases we are considering a very changeling proposal as we are having different visions to explain causality. Only if the benefits are large and the integration is clearly defined, it would be done.

Case (*b*): InfA ouput is explained as a mechanistic causal relation (such as ODE system). In this case the integration would be considering by adjusting time–scales.

However, one very important point is to define which elements in the InfA output are "close" to mechanistic model variables/parameters. This closeness would be done by the use of ontologies. In some cases the relation will be perfect; that is the annotation of elements in the database and in the model are equal. But in other cases it will be necessary to set closeness measures between model variables/parameters and InfA network´s nodes. "Ontology distance" is not a well-defined element and it will be studied within the WP3 and WP4.

Note that in most cases the "extension" of causality is not a standard and well-studied problem. Opposite to that, it is a very hard problem were few generic methodologies have been developed. We will study the possible options and then decide how to proceed based on the results obtained.

3.2 MA to InfA

A mechanistic model provides, by the use of causal relation embedded in mathematical models, the opportunity to observe some elements that are hardly observed in clinical trials or experimental settings. Therefore the use of MA models can extend any data-set if there are links allowing this to happen.

Figure 5 describes the process how this extension in the data set can be made. The first step is to establish which parameters are considered as observables in the data set. If at least one exists, then the non-linked parameters can be considered as fixed to physiological values. Then for each patient, the model can be run to compute the variables of the models (within Synergy models, the steady state is computed). Those variables are considered un-observables, and as they are computed for each one

of the patients they can be added as columns in the data set.

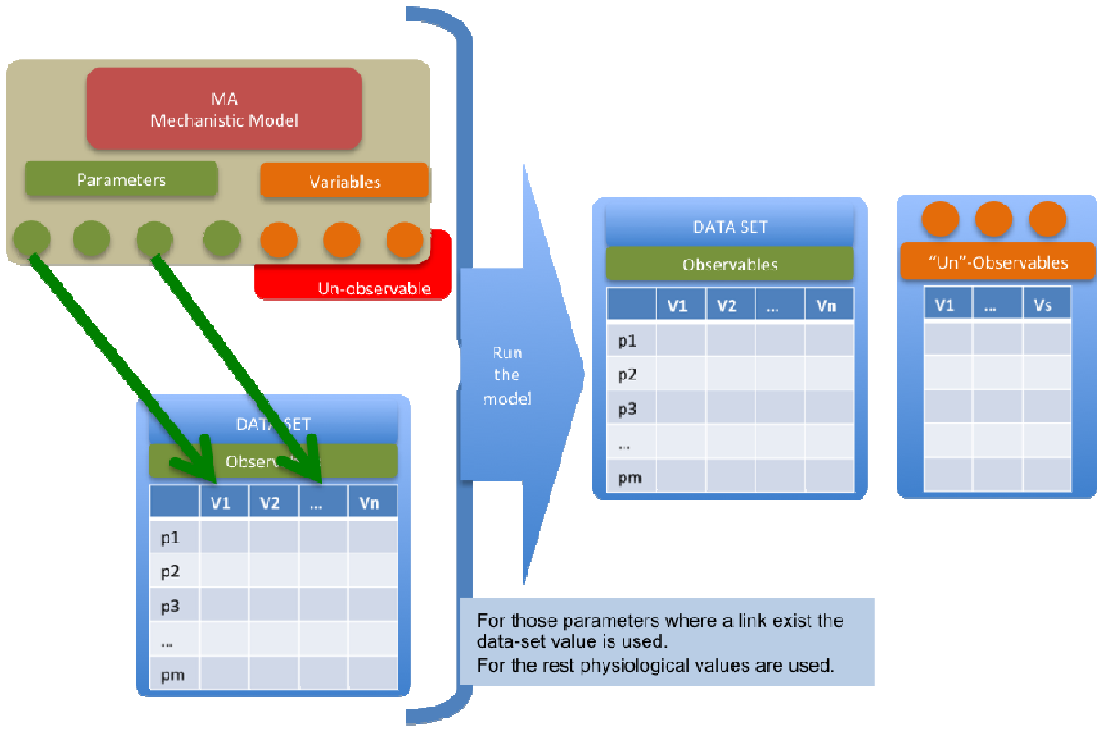


Figure 5. Extend data sets bythe use of mechanistic models.

Once the data-set is extended many different options can be considered, but majorly to "repeat" or "re-process" all the analysis done on them with the new columns. Some options are described below:

Option 1: to observe if un-observable columns can be related to other columns by simple statistical methods (such as correlations)

Option 2: to repeat the InfA network generation with the extended data set. This will provide inputs about where those new variables cluster with for instance. (See section 1.1 and Section 2.1 for more details).

Option 3: to refine clustering results. For instance PAC-COPD shows that COPD patients can be grouped in three classes with very different scenarios. Similarly to proposed in section 3.2 we can use the mechanistic models to compute non-observed molecules or phenotypes that are not in the original data set. By doing this we can improve the prediction power of the disease type predictor; if that would be the case, we would also analyze how the different variables are of relevance in class prediction.

3.3 Setting the scenario

All the methods presented here take into account different scenarios; within each scenario different proposals can be considered. However it is clear that the first main step is to set in detail which scenario are we dealing with.

Setting a scenario is understood as (1) to specify the links between data sets and models. Once this is clear, we can observe the MA to InfA approach. Also (2) it is relevant to compile in an efficient and ordered way the results provided by InfA in order to enumerate the possible "causality extension" options of interest.

4 Conclusions

The word "Synergy" stands for the outputs that can come from two or more things functioning together, that cannot be repeated by the "things" acting independently.

Following this consideration we aim in this deliverable to obtain new, non-previously observed, insights of the mechanisms of COPD progression by combining two very different sources of information: mechanistic models and Inference network approaches. To this end Section 1 first explains both sources of information separately. Section 2 specifies the methods used within Synergy-COPD constraints.

Section 3 describes the core of the deliverable, where both InfA and MM approaches are integrated in such a way the results of one are considered as inputs for the other, and viceversa. However, even if the proposal includes different working scenarios is a necessary first step to describe the actual situation.

This deliverable aim to set the stones where to build the integrative approach, therefore it becomes by definition a working document that will be updated when needed and when more information or methods are considered. Within Synergy-COPD there has been many meetings that have been used to detail the different sources of information and how to proceed in its analysis. More will come, in a bi-monthly basis to define steps in order extend the work done by now.

5 References

(1) Antczak P, Ortega F, Chipman JK, Falciani F. Mapping drug physico-chemical features to pathway activity reveals molecular networks linked to toxicity outcome.PLoS One. 2010 Aug 27;5(8):e12385.

(2) Avriel M, *Nonlinear Programming: Analysis and Methods*. 2003: Dover Publishing,.

(3) Aymerich G, Gomez J, and JM Anto, *Phenotypic Characterization and Course of Chronic Obstructive Pulmonary Disease in the Pac-Copd Study: Design and Methods.* . Archivos de Bronconeumología Archivos de Bronconeumología, 2009. **45**: p. 4-11

(4) Aymerich G et al., *(Copd) Subtypes. Identification and Prospective Validation of Clinically Relevant Chronic Obstructive Pulmonary Disease.* Thorax, 2011. **66**: p. 430-437.

(5) Bader GD, Hogue CW. An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics. 2003 Jan 13;4(1):2.

(6) Balcells E et al, *Factors Affecting the Relationship between Psychological Status and Quality of Life in Copd Patients.* Health Qual Life Outcomes, 2010. **8**(108).

(7) Balcells E et al., *Characteristics of Patients Admitted for the First Time for Copd Exacerbation.* Respiratory Medicine, 2009. **103**: p. 1293-1302

(8) Bansal M, Belcastro V, Ambesi-Impiombato A, di Bernardo D. How to infer gene networks from expression profiles. Mol Syst Biol. 2007;3:78.

(9) - , Gea J, et al. (2009) Chronic endurance exercise induces quadriceps nitrosative stress in patients with severe COPD. Thorax 64: 13–19.

(10) Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A. Reverse engineering of regulatory networks in human B cells. Nat Genet. 2005 Apr;37(4):382-90. Epub 2005 Mar 20.

(11) Bender EA, ed. *An Introduction to Mathematical Modelling*. 1978, Wiley: New York.

(12) Boyd S and Vandenberghe L, *Convex Optimization.* . 2004: Cambridge University Press.

(13) Butcher JC, *Numerical Methods for Ordinary Differential Equations* 2003.

(14) Butte AJ, Tamayo P, Slonim D, Golub TR, Kohane IS. Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. Proc Natl Acad Sci U S A. 2000 Oct 24;97(22):12182-6.

(15) Cross M and Moscardini AO, *Learning the Art of Mathematical Modelling*, ed. E.H. Ltd. 1985, Chichester.

(16) de Batlle J and et al., *Dietary Habits of Firstly Admitted Spanish Copd Patients.* Respiratory Medicine, 2009. **103**: p. 1904-1910.

(17) Ellner SP and Guckenheimer J, *Dynamic Models in Biology*. 2006: Princeton University Press.

(18) Fowkes ND and Mahony JL, *An Introduction to Mathematical Modelling,*. 1994, Chichester: John Wiley and Sons.

(19) Frederic BJ, Charles GJ, Claude L, and Claudia A, *Numerical Optimization: Theoretical and Practical Aspects*. 2006, Berlin: Springer-Verlag.

(20) Frédéric BJ and Alexander S, *Perturbation Analysis of Optimization Problems*. Springer Series in Operations Research, ed. Springer-Verlag. 2000, New York.

(21) Frey HC and Patil SR, *Identification and Review of Sensitivity Analysis Methods.* Risk Anal, 2002. **22**(3): p. 553-78.

(22) Grizzi F and Chiriva-Internati M, *The Complexity of Anatomical Systems.* Theor Biol Med Model, 2005. **2**(26).

(23) Gupta R, Stincone A, Antczak P, Durant S, Bicknell R, Bikfalvi A, Falciani F. A computational framework for gene regulatory network inference that combines multiple methods and datasets. BMC Syst Biol. 2011 Apr 13;5:52.

(24) Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. Nature Protoc. 2009;4(1):44-57

(25) Johnson, WE, Rabinovic, A, and Li, C (2007). Adjusting batch effects in microarray expression data using Empirical Bayes methods. Biostatistics 8(1):118-127.

(26) Kaplan S and Garrick BJ, *On the Quantitative Definition of Risk.* Risk Analysis, 1981. **1**(1): p. 11-27.

(27) Larson RE, Hostetler RP, and Edwards B, *Calculus with Analytic Geometry*. 1994, Lexington, MA.

(28) Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, Califano A. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics. 2006 Mar 20;7 Suppl 1:S7.

(29) Maria A. *Introduction to Modelling and Simulation*. in *Proceedings of the Winter Simulation Conference* 1997.

(30) Nocedal J and Wright SJ, *Numerical Optimization*. 2006: Springer.

(31) Nunez B et al., *Anti-Tissue Antibodies Are Related to Lung Function in Chronic Obstructive Pulmonary Disease.* Am. J. Respir. Crit. Care Med. , 2011. **183**: p. 1025-1031.

(32) Rangel C, Angus J, Ghahramani Z, Lioumi M, Sotheran E, Gaiba A, Wild DL, Falciani F. Modeling T-cell activation using gene expression profiling and state-space models. Bioinformatics. 2004 Jun 12;20(9):1361-72. Epub 2004 Feb 12.

(33) Selinger DW, *On the Complete Determination of Biological Systems.* Trends Biotechnol, 2003. **21**: p. 251–254.

(34) Selivanov VA and et al, *Bistability of Mitochondrial Respiration Underlies Paradoxical Reactive Oxygen Species Generation Induced by Anoxia.* PLoS Comput Biol, 2009. **5**.

(35) Selivanov V, Votyakova T, Pivtoraiko V, and Cascante M, *Reactive Oxygen Species Production by Forward and Reverse Electron Fluxes in the Mitochondrial Respiratory Chain.* PLoS Comput Biol, 2011. **7**(3).

(36) Stewart J, *Calculus, International Metric Edition 6e*. 2008, McMasterUniversity BrooksCole,.

(37) Szallasi Z, Stelling J, and Periwal V, *System Modeling in Cell Biology: From Concepts to Nuts and Bolts*. Vol. xiv. 2006, Cambridge, MA: MIT Press. 448.

(38) Turan N, Kalko S et al. (2011). A Systems Biology Approach Identifies Molecular Networks Defining Skeletal Muscle Abnormalities in Chronic Obstructive Pulmonary Disease. In press PLOS Comp. Biol.

(39) Werhli A, Grzegorczyk M, Husmeier D 2006 Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical Gaussian models and Bayesian networks. Bioinformatics. 22, 2523–2531

(40) Yip KY, Alexander RP, Yan K-K, Gerstein M: Improved Reconstruction of In Silico Gene Regulatory Networks by Integrating Knockout and Perturbation Data. PLoS ONE 2010, 5(1):e8121