# CNS PERTINENT BIOMARKERS RESEARCH. PROGRAM OVERVIEW


**Modeling the physiological mechanisms associated with pathological anxiety to reveal pertinent biomarkers.**
**(This approach applies to all disease domains)**

## *Biomarkers search: Context & Issues.*

Biomarkers for improved prediction and monitoring of disease and toxicology mechanisms are needed to control the high clinical failure rates among new compounds. Strong efforts are being devoted to the search for combinatorial biomarkers, generated through high content screening, and in particular high content in situ transcriptomics, proteomics and imaging technologies.

But these approaches can only identify “biomarkers” as a function of their statistical occurrence and not in terms of their physiological relevance. Hence, the maturity and utility of currently available biomarkers varies very significantly between target organ systems and most such biomarkers really reflect the nature of the lesions and seldom that of the underlying pathology.

Concurrently, the vast majority of complex disorders, and in particular neurological dysfunctions, are defined by a number of symptoms that can differ considerably between affected individuals with respect to their presence, frequency, severity, and topography. As a result of this semiologic heterogeneity, pathophysiologically different forms of a disease much too frequently fail to be recognized as such while heterogeneous presentations of a same pathology lead to differential diagnosis. This, inescapably, leads to sample misclassifications that can reach very significant proportions. Furthermore, individuals affected by a severe disease often present a variety of concurrently induced/associated disorders, some of which may remain under-diagnosed and their prevalence under-rated. If it is accepted that a pathology must necessarily leave traces of its presence under the form of biomarkers, then the concurrent presence of another pathology, whether clinically recognized or not, must also necessarily do so.

What does all this do to the problem of searching for co-occurrences between biological components, including proteins, on the basis of serendipity (the only possibility in the absence of pathophysiological understanding)? To have the least chance of success, knowledge of what to search for, where, when and why appears to be a necessity.

Acquisition of the necessary knowledge can be obtained, in parts, using in-silico theoretical models produced through analytical approaches and processes collectively known as “Systems Biology”.

***But the mechanisms of life being integrative and non-linear processes, they require “specific systems biology” approaches to be correctly understood.***

***The CADI™ (Computer Assisted Deductive Integration) models building approach, based on the disruptive "negative selection" concept, and an innovative synergic collaboration between experimental biology and integrative biology, has proven to allow researchers to take the best possible decisions for the best possible results in a minimum of time and resources.***

# A CNS "pertinent" biomarkers discovery program.

***The objective of this program is to qualitatively define the nature and sequence of functional, intra CNS events and mechanistic modifications that lead to the development of potentially pathological anxiety within a relatively homogenous physiological context (inbred mice in which pathological anxiety is inducible).***

The experimental data produced by the partners (see part B in CADI™ figure; gene expression and proteins dynamics profiles) using the inducible animal model are used to qualitatively model the physiological mechanisms that best fit the experimental observations. The heuristic, non-mathematical models produced by the approach we utilize allow to define and classify events and mechanisms in terms of both their origins (astrocytes, microglia, endothelium, podocytes, neurons, etc.) and their functional consequences.

To this end, the expression profiles (genes & proteins) are mined to build hypothetical biological mechanisms. We then extract prediction of co-expression patterns that should be associated with these mechanisms if they are active/inhibited. Any hypothesis that can be refuted using the data sets and/or the literature (context taken into account) is abandoned (negative selection process). The hypotheses that resisted destruction attempts often include genes that are not in the data set, but at this stage, that is irrelevant. These hypothetical mechanisms are then integrated to produce models from which either physiological or physical behaviors can be extracted. These are in turn confronted to the data sets and the literature and subjected to the negative selection procedure. The process is iteratively repeated until a model that cannot be refuted is obtained. This model must account for the genes/proteins patterns predicted by the hypotheses that are now included in it, but were absent from the data set.
This model predicts very precise components behaviors and mechanisms (activated and inhibited), not obviously represented in the data sets but that can be directly verified by experimentation in vivo.
In general, the first version of such a model contains many errors. The new data, generated by the experiments designed to test the model, is in turn integrated as above, thereby correcting the model. In our experience, after two or three such rounds, the model clearly depicts in details the functional physiological basis of the phenomenon studied.
This constitutes a road map. And since this map indicates the key events and mechanisms associated with the phenomenon, it becomes possible to address the question of consistency. Now knowing what to investigate, where, when and why, heterogeneous data sets (different mice strains) can be fruitfully utilized. Co-expression/regulation networks that depict the situation validated in vivo are constructed and the heterogeneous data sets analyzed to identify what correlates with and what differs from the initial situation. This analysis allows to define the core characteristics associated with the various forms of the phenomenon, albeit in terms of physiological mechanisms and not merely in terms of patterns. This directly leads to the isolation of pathophysiologically relevant biomarkers, potential therapeutic approaches, etc.

## *Non-BMSystems' proprietary data management & analytical systems used in such program.*

***Most of the data management and analytical softwares and systems utilized throughout such a program are web-based open source systems supporting the major standards of data representation and exchange*** (MAGE-ML, DiGIR, XML, RDF, etc.) while providing scalability (SOA/OGSAI, MapReduce and OGSA-DQP). BMSystems utilizes a large set of such softwares.

## *BMSystems' proprietary Integrative tools and methodology.*

***The main model construction tasks are carried out using the methodologies and tools proprietary to BMSystems. CADI™ Architect*** allows to contextualize, compile and structure various forms of information (texts, images, charts, etc) within a common environment. ***CADI™ Search*** mines this information within the context of functional hypotheses. It allows to determine whether or not an hypothesis can be refuted using accessible information and if so to what extent and by which types of information. ***Fiber-N™*** is the integrative engine that allows to link individual corroborated functional hypotheses into hypothetical physiological mechanisms and to structure them under the form of metahypotheses to be tested using *CADI Search.* ***CADI™ View*** allows to graphically depict, in various levels of details, the components that constitute hypothetical physiological mechanisms under format compatible with freewares such as *Cytoscape* and *CellDesigner*.

Such a program generates consistent proteomics and transcriptomics datasets that must be made available to all participants. The DAVID and Bioconductor packages provide flexible environments to manage and analyze such data sets. The principal investigator in each team is responsible for smooth data transfer to a web-accessible central server running an MySQL database together with the *DAVID*, *C-Path*, *Cytoscape* and *CellDesigner* softwares, allowing the partners to load, access, manipulate and share data at all processing stages.

## *BMSystems competitive advantages vs. current competition*

What distinguishes this BMSystems program is the innovative, results oriented systems biology approach followed.

Most current systems biology projects utilize gene co-expression networks and dynamic correlations approaches aiming to extract, from heterogeneous data sets representing various phenotypic presentations of a given biological phenomenon, genes-proteins signatures that could functionally characterize the phenomenon studied. This is followed by mathematical modeling of the corresponding pathways, generally under the form of logical (Boolean) networks, and confrontation of the model with data obtained from the relevant biological system.

***While current approaches have been successfully applied to some in vitro systems based on the utilization of cell lines, they have so far largely failed when addressed to problems involving tissues or physiological interactions between different cell-types.***

## *There are at least three different reasons for this:*

1) The logic behind gene co-expression networks and dynamic correlations approaches is really that of a "social network" that assumes linearity in relationships: functionally related genes could be clustered together based on expression profiles. In effects, the principle assumes that component dynamics directly reflects or predicts functional characteristics.

That is unfortunately not the case. Biological systems are surprisingly noisy and leaky. However, the noises and leaks are far from being entirely stochastic as is clearly assumed by the above approaches. They allow long and short-range modulations within networks and that is a key aspect of biological systems. As a result, networks components dynamics show "apparently dissociated" fluctuations that are much wider then expected and generally not accommodated by the limits that must be imposed to distinguish a "co-regulated network" from a "randomized" one. This results in the production of clusters that are either much too large or functionally much too heterogeneous to be transformed into testable physiological mechanisms.
When applied to a tissue as functionally complex as the CNS, such approaches cannot be expected to produce fruitful results.

2) Logical networks solely rely on the known network structures of a pathway. In effect, they define, in terms of pathways & hubs, what is known to exist and not what is contextually coherent.

Indeed, while relying heavily upon massive literature analyses, this model-building approach does not take into account the facts that the published reports utilized are always incomplete (to an unknown extent), biased (in unknown manners and to an unknown extent) and erroneous (to an unknown extent). In the results subsequently delivered the "true" is mixed with the "uncertain" without possibilities to eliminate the "false" and even less to determine in what contexts the "true" may suddenly become "false". In circumstances where the contributions of the different cell-types that constitute a tissue must be clearly distinguished and taken into account, the resulting massive levels of inconsistencies generated by the numerous conflicting or physiologically incompatible aspects that have been included and not detected inescapably lead to analytical failure.
Hence, the higher the level of internal heterogeneity introduce in the experimental system, the worse the analytical situation gets.

3) Heterogeneous data sets (different phenotypes) seldom do correspond to different points along a continuum (successive transitions in state) but frequently reflect differences in physiological fate (divergent paths) arising from in-built physiological deviations associated with each phenotype.

When knowledge concerning the physiological basis of the phenomenon studied is available, such data sets become highly useful indeed since they can be utilized to test its degree of consistency. However, their utilization to uncover this very physiological basis in the first place is highly ineffective since individual data set, most likely reflecting entirely different physiological contexts, cannot bear direct comparisons.

The approach we are following, while certainly iterative (the interplay between modeling and experimentation is indeed absolutely essential) and applicable to both in vitro and in vivo biological systems, is not dependent upon high throughput or high contents analysis.
It is, however, highly dependent upon comparative analyses.
This implies that the experimentations must be carried out on conditional or inducible animal model and not upon "constitutive" models. In such a context, the subjects (induced) and controls (non-induced) having highly similar genetic and physiological backgrounds, the differences in gene expression and proteome dynamics between controls and affected become extremely useful because they correspond to an actual transition in state. The problem can thus be analytically approach under the form of a continuum, closely reflecting clinical reality.
The changes in expression profiles (genes and proteins) can now be used for modeling purposes.

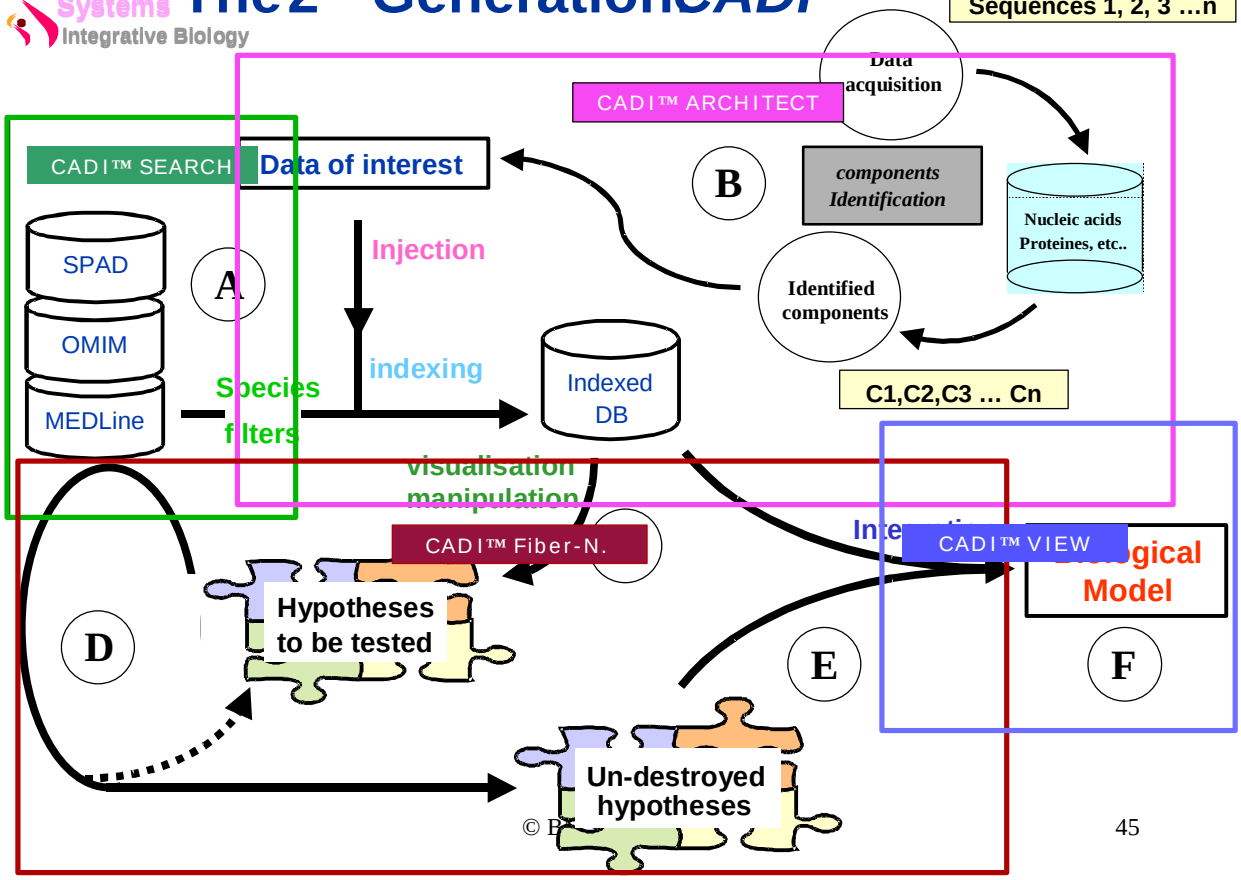
***Furthermore, the logic behind the model-building approaches we use does not assume linearity and the components of a model do not incorporate solely what is known. Our approaches rely upon strict negative selection processes. Hence a model arising from this procedure contains elements that have never yet been described but cannot be refuted by current knowledge and/or available biological data. As a result, the models of complex human diseases we have so far constructed using this approach have led to discoveries that became the objects of patent protections in both the pharmacological and biotechnological domains.***

Dr. François iris; Founder & CSO BMSystems

For more information

Manuel GEA
Co-founder & CEO
BIO-MODELING SYSTEMS
Predictive Integrative Biology
26 rue Saint Lambert
75015 PARIS FRANCE
+33 6 83 06 12 72
email: manuel.gea@bmsystems.net
www.bmsystems.net

# The 2nd Generation *CADI™*






# BMSystems outputs

***What can we do with CADI™ models?***

***Reduce time to result, improve success rate and reduce costs to exploit:***