How scientific literature analysis yields innovative therapeutic hypothesis through integrative iterations
Francois Iris, Athanasios Beopoulos and Manuel Gea

It is becoming generally accepted that the current diagnostic system often guarantees, rather than diminishes, disease heterogeneity. In effects, syndrome-dominated conceptual thinking has become a barrier to understanding the biological causes of complex, multifactorial diseases characterized by clinical and therapeutic heterogeneity. Furthermore, not only is the flood of currently available medical and biological information highly heterogeneous, it is also often conflicting. Together with the entire absence of functional models of pathogenesis and pathological evolution of complex diseases, this leads to a situation where illness activity cannot be coherently approached and where therapeutic developments become highly problematic. Acquisition of the necessary knowledge can be obtained, in parts, using in silico models produced through analytical approaches and processes collectively known as ‘Systems Biology’. However, without analytical approaches that specifically incorporate the facts that all that is called ‘information’ is not necessarily useful nor utilisable and that all information should be considered as a priori suspect, modelling attempts will fail because of the much too numerous conflicting and, although correct in molecular terms, physiologically invalid reports. In the present essay, we suggest means whereby this body of problems could be functionally attacked and describe new analytical approaches that have demonstrated their efficacy in alleviating these difficulties.

Address
Bio-Modeling Systems, Tour CIT, 3 Rue de l’Arrivée, 75015, Paris, France

Corresponding author: Iris, Francois (francois.iris@bmsystems.org)

How scientific literature analysis yields innovative therapeutic hypothesis through integrative iterations
Francois Iris, Athanasios Beopoulos and Manuel Gea

It is becoming generally accepted that the current diagnostic system often guarantees, rather than diminishes, disease heterogeneity. In effects, syndrome-dominated conceptual thinking has become a barrier to understanding the biological causes of complex, multifactorial diseases characterized by clinical and therapeutic heterogeneity. Furthermore, not only is the flood of currently available medical and biological information highly heterogeneous, it is also often conflicting. Together with the entire absence of functional models of pathogenesis and pathological evolution of complex diseases, this leads to a situation where illness activity cannot be coherently approached and where therapeutic developments become highly problematic. Acquisition of the necessary knowledge can be obtained, in parts, using in silico models produced through analytical approaches and processes collectively known as ‘Systems Biology’. However, without analytical approaches that specifically incorporate the facts that all that is called ‘information’ is not necessarily useful nor utilisable and that all information should be considered as a priori suspect, modelling attempts will fail because of the much too numerous conflicting and, although correct in molecular terms, physiologically invalid reports. In the present essay, we suggest means whereby this body of problems could be functionally attacked and describe new analytical approaches that have demonstrated their efficacy in alleviating these difficulties.

Address
Bio-Modeling Systems, Tour CIT, 3 Rue de l’Arrivée, 75015, Paris, France

Corresponding author: Iris, Francois (francois.iris@bmsystems.org)

How scientific literature analysis yields innovative therapeutic hypothesis through integrative iterations
Francois Iris, Athanasios Beopoulos and Manuel Gea

It is becoming generally accepted that the current diagnostic system often guarantees, rather than diminishes, disease heterogeneity. In effects, syndrome-dominated conceptual thinking has become a barrier to understanding the biological causes of complex, multifactorial diseases characterized by clinical and therapeutic heterogeneity. Furthermore, not only is the flood of currently available medical and biological information highly heterogeneous, it is also often conflicting. Together with the entire absence of functional models of pathogenesis and pathological evolution of complex diseases, this leads to a situation where illness activity cannot be coherently approached and where therapeutic developments become highly problematic. Acquisition of the necessary knowledge can be obtained, in parts, using in silico models produced through analytical approaches and processes collectively known as ‘Systems Biology’. However, without analytical approaches that specifically incorporate the facts that all that is called ‘information’ is not necessarily useful nor utilisable and that all information should be considered as a priori suspect, modelling attempts will fail because of the much too numerous conflicting and, although correct in molecular terms, physiologically invalid reports. In the present essay, we suggest means whereby this body of problems could be functionally attacked and describe new analytical approaches that have demonstrated their efficacy in alleviating these difficulties.

Address
Bio-Modeling Systems, Tour CIT, 3 Rue de l’Arrivée, 75015, Paris, France

Corresponding author: Iris, Francois (francois.iris@bmsystems.org)

Background
Drug development is primarily a problem of data integration and knowledge management. Knowing the potential targets of a molecule and the functions of these targets is one thing. Understanding the physiological mechanisms that must be targeted and the manner in which they must be manipulated to have a therapeutic impact is quite another. Thus, success in therapeutic development largely depends upon the coherent manipulation of a physiological system in its pathological context and not upon the manipulation of a target in a molecular setting. Yet, identification of the presence of a given pathology is largely based upon the symptoms presented by any given patient. These symptoms, together with the results of medical and biological tests, are then utilised to reach a medical diagnostic. In practice, most experienced physicians utilise the pattern recognition method to identify the clinical problem. Theoretically, a given pattern of tests results and symptoms within a given local population context can be directly associated with a given therapy, even without a definite decision regarding what is the actual disease [1].

Hence, the vast majority of complex disorders are defined by a number of symptoms that can differ considerably between affected individuals with respect to their presence, frequency, severity and topology. Indeed, within a population context, different individuals may present similar symptoms for totally different physiological reasons just as they can present different symptoms for very similar physiological reasons. However, the compromise that constitutes the pattern recognition method, which primarily relies upon the information available to the physician, carries a substantial risk of misdiagnosis, confusing different pathologies which actually require different therapies. This is most evident in the context of complex pathologies [2–5].

Furthermore, heterogeneity in symptoms complicates the search for the aetiology of complex diseases and the mechanisms for their treatment. In effects, the current diagnostic system often guarantees, rather than diminishes, disease heterogeneity and current syndrome-dominated conceptual thinking has become a barrier to understanding the biological causes of a wide variety of diseases characterized by clinical and therapeutic heterogeneity such as muscular dystrophies [6], mitochondrial dysfunctions [7,8], retinal degenerative diseases [9,10], thyroid pathologies [11], autoimmune [12] and neurological diseases [13,14], metabolic [15**,16] and psychiatric disorders [17], and so on.

This leads to an untenable situation that precludes coherent therapeutic developments since it effectively
The issues of biomarkers in drug development

Biomarkers are at the roots of evidence-based medicine (who should be treated, how and with what) and without valid biomarkers, not only advances in better targeted therapies will remain limited but treatments will also remain largely empirical. Furthermore, biomarkers for improved prediction and monitoring of disease and toxicology mechanisms are needed to control the high clinical failure rates among new compounds [18,19].

But, in the absence of clear pathophysiological understanding, the maturity and utility of safety-related biomarkers varies very significantly among target organ systems [20*,21].

A ‘biomarker’ is typically defined as a laboratory measurement that reflects the activity of a disease process or the responses to a therapeutic intervention. But the goals of therapeutic interventions are twofold: (1) better symptomatic therapies, and (2) treatments that slow disease progression or delay disease onset. This necessarily leads to a second class of biomarkers, known as ‘clinical endpoints’, that are not measured for the purpose of detecting clinical benefit but for their reflection of the underlying pathological process [22]. In essentially all cases, these markers must quantitatively correlate, either directly or inversely, with disease progression. Taking into account the state of our current understanding of pathological processes, this literally opens a Pandora box. In the context of functionally heterogeneous disorders, there might be as many biomarkers as there are affected individuals. Hence, the much sought-after ‘gold standard biomarkers’ for a set of individuals affected by a common disease remains an unattainable goal [23].

In attempts to circumvent these issues, a third class of biomarkers has been put forward, the so-called ‘surrogate markers’. This object is defined as a laboratory measurement or physical sign that is used in therapeutic trials as a substitute for a clinically meaningful endpoint and is a direct measure of how a patient feels, functions, or survives and is thus expected to predict the effects of a therapy [24]. Hence, the major difference between a biomarker and a surrogate marker is that a biomarker is a ‘candidate’ surrogate marker, whereas a surrogate marker is a test used, and taken, as a measure of the effects of a specific treatment.

However, drugs development must necessarily proceed through pre-clinical studies carried out on laboratory animal models, usually inbred rodent strains, which present the apparent symptomatology of the human pathology being addressed but rarely its actual physiological basis. Not only these animal models often amount to mere caricatures of the human pathology, but the results of the drugs development assays are also interpreted according to their effects upon the animal model’s symptomatology, hence an all too frequent inadequacy with respect to the human physiopathology with ensuing clinical trial failures or drug withdrawals from the market.

As a result, strong efforts are now being devoted to the search for combinatorial biomarkers, generated through high content screening, and in particular high content in situ proteomics and imaging technologies, to be used in the industry to screen for toxic side effects of drug candidates and to identify appropriate patient populations [25] in the hope that this will support the knowledge-based decision-making process by providing crucial information on functional biology [26,27]. In doing so, it is assumed that a given symptomatically defined disorder (semiology) necessarily implicates restricted sets of physiological mechanisms, some of which must eventually be shared among affected patients, irrespective of their environments. It thus becomes a matter of screening a sufficiently large number of patients to hence identify relevant markers or combinations thereof.

These approaches therefore identify ‘biomarkers’ as a function of their statistical occurrence and not in terms of their physiological relevance. The phenomena that give rise to disease and responses to treatments heterogeneities, the very reason behind the search in the first place, are entirely ignored. Furthermore, individuals affected by a severe disease often present a variety of concurrently induced/associated disorders (comorbidities), some of which may remain under-diagnosed and their prevalence under-rated [28*,29,30]. If it is accepted that a pathology must necessarily leave traces of its presence under the form of biomarkers as defined above, then the concurrent presence of another pathology, whether clinically recognised or not, must also necessarily do so.

Thus, far from helping to resolve the issues generated by the syndrome-dominated vision, this further worsens an already difficult situation, particularly in the case of heterogeneous disorders. These shortcomings have for net results to reiterate previous costly mistakes, albeit under a different form. Not only statistical effects are expected to compensate for lack of knowledge, but an additional flaw is now being introduced. Differences in physical environments are implicitly considered as having little impact upon the biological mechanisms associated with defined semiotics [31*,32]. This directly leads to a highly deleterious situation already experienced by the industry in the past.

Indeed, in order to increase drug development successes, it was found necessary to significantly increase the size and the scope of clinical trials. The main reasons for this were associated with the phenomenon of functional
disease heterogeneity coupled with the impossibility to define trustworthy exclusion criteria applicable at recruitment levels [33]. The results are a massive escalation in costs with very limited returns in terms of successes [34].

**The biological reality**

Semiology is a poor indicator of similarities or differences in associated physiological processes. Not only functionally different disease states can present similar semiotics, but a disease with well identified biological causes often present considerable variability in both semiology and outcome among affected patients, including between siblings [35,36]. Here, both physical and sociological environments play important roles [37–39]. All the more so since they also affect epigenetic mechanisms and subsequent disease susceptibility without change in primary DNA sequence [40,41,42**]. All these effects are particularly apparent in the numerous cases of discordances among monozygotic twins for a wide variety of disorders, many of which are regarded as having significant genetic backgrounds [43,44].

The outcome is that physiopathologically different forms of a disease much too frequently fail to be recognised as such while heterogeneous presentations of a same physiopathology lead to differential diagnosis [45,46]. This, inescapably, leads to sample misclassifications that can reach very significant proportions [47]. Hence, given the biological reality briefly described above, it seems hardly reasonable to expect that useful potential clinical and/or prognostic and/or therapeutic biomarkers could be identified on the basis of their statistical occurrence. To have the least chance of success, knowledge of what to search for, where, when and why appears to be a necessity.

**Biological functions results from interactions between integrative and non-linear mechanisms**

Phenotypes and behaviour depend on the integrated effects of multiple signalling pathways and molecules, genetic polymorphisms, epigenetic marking dynamics and epistatic phenomena as well as environmental stimuli affecting post-translational processing. This is true for all biological systems, from individual cells all the way to organisms.

Interventions upon such systems must take into account the fact that biological functions result from dynamic, integrative and non-linear processes subject to discontinuities. Indeed, it has become obvious that the mere state assessment of components and metabolites in a living system does not reveal in a predictable and reliable manner the activity of pathways and circuits that they comprise [48,49]. Furthermore, many organs and biological functions are eminently affected by time (the act of being actively alive), in terms of both structural anatomy as well as networks interactions [50,51]. Functional evolution over time (ageing) implies numerous switches in gene expression patterns [52,53]. Interpreted in terms of functional physiology, this leads to the inescapable conclusion that, in any individual, the functional situation implemented when time-point ‘B’ will be reached may be radically different from what it was at time-point ‘A’. The numerous reported cases of spontaneous remission in a variety of sever pathological conditions [54–57] may well be striking examples of this.

Hence, it is not merely a matter of what proteins (components) are expressed and to what level, but rather of what other potential interaction partners are present and in what state. It is qualitative aspects that are important here, and not quantitative considerations. A protein deemed ‘physiologically important’ may be entirely absent or constitutively non-functional without producing deleterious phenotypic effects. This is amply demonstrated by the numerous such instances observed in knockout mice [58]. Furthermore, the deleterious effects of inactivating mutations affecting a given protein can often be compensated by inactivating mutations simultaneously affecting another protein [59] or by corrective mechanisms its functional absence induce [60**,61,62].

**The needs for systemic biological models and the role of systems biology**

The necessarily brief survey above provides some indication of the density, extent and enormous complexity of the integrative effects, resulting from dynamic interactions between a multitude of biological factors and interconnected pathways, leading to most medical conditions requiring therapeutic interventions [63**,64]. This makes simplistic pharmacological approaches ineffective when not actually damaging [65,66]. Indeed, in most instances, pharmacological interventions take the form of monotherapies, each targeting a very limited, if not a single set of physiological mechanisms. Besides the numerous undesirable effects these approaches trigger, which are often at the root of pre-clinical and, more damagingly, clinical failures [67*], de novo and acquired resistances to treatment are widespread [68–70]. It is becoming evident that, for the vast majority of complex disorders, therapeutic approaches that simultaneously target multiple pathways are urgently required [71–73].

The challenge for future drug development will thus be to devise pharmacological approaches that reflect the overall pathophysiological and biological processes that must be affected to produce defined therapeutic responses whilst retaining the simplicity and robustness required for routine pre-clinical and clinical testing.

Appropriate pathophysiological understanding will therefore be necessarily required for both therapeutic targeting purposes as well as for the identification of biomarkers that are physiologically relevant for clinical (patients
stratification), prognostic (disease status) and therapeutic (response/efficacy/toxicity) purposes [74**].

The analytical and biological complexities to be mastered are such that none of this can be achieved using reductionist (classical) approaches. To have the least chance of success, we must develop predictive functional models sufficiently detailed so as to enable the precise identification, in mechanistic terms, of events leading to pathological consequences, hence identifying the mechanisms and the markers associated with these events together with the modes of intervention most likely to prevent or alleviate the problems. 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’. Systems biology addresses the need to shift from a component-based reductionist view of biology to a system-wide perspective. It can be described as a global analysis of how all components in a biological system interact to determine its phenotype. Although the definitions may vary, systems biology can usually be characterized as interdisciplinary, iterative, computationally intensive, and information greedy.

**Systems biology applied to the discovery and validation of novel therapeutic approaches**

The usual approaches to systems biology are characterized by

1. their dependence upon relevant quantitative data arising from multiple targeted experimental interrogations in an iterative interplay between experimentation and modelling, with
2. the aim of elucidating how the molecular components of a living system determine its phenotype by exploring their dynamic interplays as well as their interactions with the environment.

Thus, this approach interprets biological phenomena as dynamic processes, the mechanisms and consequences of which depend on the behaviour of components that constitute the living entity studied.

While this requirement may be reasonably fulfilled in the context of well-defined mechanisms (e.g. the Rab5–Rab7 toggle and cut-out switches in the conversion of early endosomes into late endosomes) [75], it can hardly be contemplated when addressing complex, heterogeneous and often obscure pathophysiological mechanisms. Indeed, when dealing with issues addressing complex human pathologies, data on relevant molecular and intercellular dynamics is seldom available to an extent and a range likely to sustain classical modelling approaches [76**]. Here, rather than adamantly adhering to the established principles attached to the most frequently used approaches to systems biology, attempts could be made to reconsider the problem under a different light.

One such alternative approach to systems analysis could be based on the exploitation of the tremendous amount of information contained in the existing published scientific literature and in highly heterogeneous publicly accessible databases rather than on homogeneous datasets arising from specifically targeted investigations.

However, the matter is much more arduous than might be anticipated.

**‘Information’ is a double-edged tool to be manipulated with caution**

It is indeed certain that, if predictive models of complex and heterogeneous diseases are to ever be constructed, enormous masses of information originating from a multitude of biological investigations and encompassing an enormous functional complexity will have to be integrated.

Although daunting in amplitude, if approached coherently, the flood of highly heterogeneous, and often conflicting, information that currently hampers most biological fields can become an invaluable tool. But this tool must be approached and manipulated with extreme caution.

Most of the huge amounts of currently available information arose from reductionist approaches that, in attempts to compensate for the enormous, and often insurmountable, experimental difficulties presented by in vivo systems, utilised in vitro experimentations on material frequently far removed from functional physiological reality. Hence, the enormous diversity of information obtained in association with most physiological networks represents a highly distorted, compounded view of the various modulations that can potentially affect each such network, albeit without distinction of actual in vivo physiological relevance. Furthermore, due to the wide diversity of biological systems that gave rise to this information, and the often complete lack of physiological compatibilities between experimental systems, the information thereby generated is necessarily always (1) incomplete, to an unknown extent; (2) biased, in unknown manners and to an unknown extent; and (3) erroneous, to an unknown extent, and this irrespective of the domain addressed. Indeed, it is currently considered that an estimated 85% of research resources are wasted, many published research findings being false or exaggerated [77–79,80***].

As a consequence, without analytical approaches that specifically incorporate the facts that all that is called ‘information’ is not necessarily useful nor utilisable and that all information should be considered as a priori
suspect, modelling attempts will fail because the much too numerous conflicting and physiologically invalid reports will inevitably lead to an accumulation of analytically crippling inconsistencies.

However, the above very real difficulties certainly do not mean that the scientific literature cannot be used for the implementation of systems-based analyses. These difficulties simply highlight the fact that to coherently utilise the highly heterogeneous information available, novel analytical approaches become necessary.

**Changing the analytical paradigm**

By focusing mainly on chemical and physical processes with the expectation that living systems can be fully explained from this engineer’s perspective, the classical approach to systems biology assumes bottom-up causation, from molecular dynamics to cellular/tissue behaviour.

However, the stability of a living system lies in its homeostatic capacity to re-establish itself.

In a living system, the outcome does not crucially depend on strictly predefined operations of the parts. Rather, the structure of the whole determines the operation of the parts. Indeed, almost all homeostatic processes are complex context-dependent entities to which genes make a necessary, but only partial, contribution.

In such systems, homeostasis proceeds on the basis of functional loops wherein on-going events tell local physiological contexts how to evolve, contexts tell components how to behave and components tell future events how to arise, and so on.

In other words, specific biological events do not occur because they are fated to. They occur because other events could not arise.

It follows that analyses in terms of biological components and functions now become irrelevant. What become necessary are event-driven analytical approaches.

In addition to this, the intrinsic value of any ‘information’ is only relative. It can be profoundly modified by other, indirectly linked information as well as by the contexts it can be attached to.

Thus, both the available information and the biological processes to be considered are characterized by heavily context-dependent attributes.

Therefore, whatever event-driven analytical approach is implemented, it must also be ‘relativistic’. That is, all available information must be treated on the basis of a negative selection procedure. What can be identified as false can in turn be used to discover what could be true, provided that a heuristic and event-driven analytical procedure is implemented.

Heuristics can be characterized as a problem solving approach evaluating each step in a process, searching for satisfactory solutions rather than for optimal solutions, using all available qualitative information. Thus, heuristic modelling starts from accumulated knowledge to produce, via iterative integrations, a model capable of describing the biological events and the mechanisms that generated the observed biological phenomenon and predict the modifications they will sustain in association with any given intervention.

The above considerations constitute the functional basis on which the CADI (computer-assisted deductive integration) analytical procedure was developed.

The logic behind this model-building approach (Figure 1) does not assume functional linearity and the components of a model do not incorporate solely what is known. Indeed, since this approach relies upon strict and systematic implementation of negative selection of hypotheses, models arising from this procedure contain elements that had hitherto never been described but cannot be refuted by current knowledge and/or available biological data, thereby generating novel understanding.

Here, heuristic and mathematical modelling, far from being antagonistic, are complementary. Heuristic modelling plays the role of an architect (defines the nature, the structure, the functionalities and the contextual constraints of the system under study) whereas mathematical modelling, to be implemented at a later stage, plays the role of an engineer (reveals the dynamics and robustness of the structures within the system while defining the set of parameters sufficient to give rise to similar or very different phenotypes).

Although the models arising from this analytical approach cannot, by any means, be regarded as biologically true in the absolute, they do represent a ‘least biased’ and detailed view of the mechanisms potentially associated with a given physiological state and/or governed by the biological components under consideration, with precise indications of the means whereby these could be manipulated together with the identification of the most relevant biomarkers and their significance.

In other words, these models clearly indicate what should be biologically observed in a given context, where, when, how and why. Data newly gathered from investigations aiming to test the model’s validity can then be re-injected in the model-building procedure, allowing rapid and efficient correction of the model, thereby rapidly leading to a clear and factual understanding of the biological
processes under investigation while now providing a therapeutically directly exploitable representation of the biological reality addressed.

**Conclusion**

The models produced through this approach belong to the class of non-mathematical, holistic and heuristic models. By abandoning the usual symptomatic point of view, they perform the necessary switch to diseases analyses from a causal standpoint. They constitute detailed maps of inter-cellular and/or intra-cellular mechanisms associated with a biological state and they describe the progression mechanisms of a pathological process as well as the functional characteristics of pathological state versus control.

This alternative model-building approach, which synergistically associates algorithmics and heuristics, results in a situation where biological investigations proceed at a rapid pace (months instead of years) with a hitherto unachievable success rate.

Indeed, this approach has repeatedly proven its efficacy in the discovery of (i) hitherto unsuspected biological mechanisms, pathways and interactions directly associated with phenotypic transitions *in vivo* (be they pathological or...
Moreover, when applied to neurodegenerative disorders, this approach was selected by the EU’s DG Research as one of three examples of ‘state-of-the-art’ in systems biology that benefit to medicine [90] and was granted an industrial “Best Practice Award” by The Cambridge Health Tech Institute (USA) [91].

However, it is important to realize that such models can only be an approximation of biological reality. Furthermore, the more complex the reality attached to a model, the coarser the model will be. It is therefore indispensable that such a model be confronted to the biological reality it claims to represent.

Conflicts of interest statement
Nothing declared.

References and recommended reading
Papers of particular interest, published within the period of review, have been highlighted as follows:

- of special interest
- of outstanding interest


Gives a very good account of the reasons why syndrome-dominated thinking impedes understanding diseases characterized by clinical and therapeutic heterogeneity.


Important for understanding the needs for effective biomarkers and the difficulties attached.


Gives a good example of asymptomatic and under-diagnosed comorbidity with significant physiological consequences.


A good account of the lack of correlation between statistical occurrence and physiological relevance for biomarkers in multiples disorders.


Good analysis of the reasons behind the current inefficient of drugs development.


Good analysis of the influence of epigenetic mechanisms upon disease susceptibility, clinical progression and effects of treatment.


Important to understand the functional role of dynamic processes, including protein kinetics, in disease pathogenesis and clinical progression.


Good example of the corrective mechanisms induced by the constitutive absence of a protein deemed ‘physiologically important’.


Good account of the extent and complexity of the integrative effects, resulting from dynamic interactions between a multitude of biological factors and interconnected pathways, leading to medical conditions requiring therapeutic interventions.


Reviews the underlying reasons for discrepancies between theoretical expectations and physiological outcomes leading to drugs clinical failures.


83. Iris F., Gea M, Lamp P-H, Quereux B, Quereux Bernard: • Heuristic Modelling Applied to Epidemic Homeostasis. Computational Biophysics of the Skin. Pan Stanford Publishing; 2014:461-524. Overview of the difficulties to be faced, the means whereby they are overcome together with a concrete example addressing the elucidation of an hitherto obscure differential mechanism.


