BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 12 2010, pages 1542–1547
doi:10.1093/bioinformatics/btq169

Systems biology

Advance Access publication May 6, 2010

Computing biological functions using BioΨ, a formal description of biological processes based on elementary bricks of actions

Sabine Pérès*, Liza Felicori, Stéphanie Rialle, Elodie Jobard and Franck Molina*

Sysdiag CNRS Bio-Rad UMR 3145, Cap delta/Parc euromédecine,
1682 rue de la Valsière CS 61003, 34184 Montpellier Cedex 4, France

Associate Editor: Trey Ideker

ABSTRACT

Motivation: In the available databases, biological processes are described from molecular and cellular points of view, but these descriptions are represented with text annotations that make it difficult to handle them for computation. Consequently, there is an obvious need for formal descriptions of biological processes.

Results: We present a formalism that uses the BioΨ concepts to model biological processes from molecular details to networks. This computational approach, based on elementary bricks of actions, allows us to calculate on biological functions (e.g. process comparison, mapping structure–function relationships, etc.). We illustrate its application with two examples: the functional comparison of proteases and the functional description of the glycolysis network. This computational approach is compatible with detailed biological knowledge and can be applied to different kinds of systems of simulation.

Availability: www.sysdiag.cnrs.fr/publications/supplementary-materials/BioPsi_Manager/

Contact: sabine.peres@sysdiag.cnrs.fr; franck.molina@sysdiag.cnrs.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on December 7, 2009; revised on March 12, 2010; accepted on April 13, 2010

1 INTRODUCTION

The identification of the biological functions of all genes and gene products and the understanding of how they interact to yield a living cell is still a challenge for post-genomic biology despite the fact that a huge amount of detailed information on molecular functions has been accumulated by biologists. Indeed, such detailed functional knowledge is poorly accessible for biocomputing since it is not (or rarely) formalized and it is often partial. On the other hand, the biological systems' modelling community has developed robust formalisms, in which molecular processes can be implemented in a very simplified way. The notion of biological function of a given molecule is very complex because it can vary depending on the environment. In addition, different meanings can be given to this concept based on the level of abstraction chosen by the user (molecular, cellular, etc.). Consequently, the functions of gene products have been described in many different ways over the years (Finn *et al.*, 2008; Karp, 2000; Rison *et al.*, 2000). There is, however, a need to establish and use a common reference for functional annotations. Current databases describe processes from a molecular and cellular point of view but they use text annotations, which make it difficult to handle them. Thus, to calculate with functional annotations, one would require a formal description of such biological processes. In this context, we use the term 'to calculate' to describe the action of computing with biological processes as we do with sequences (i.e. to do process comparisons, evolution, additions, etc. or to simulate dynamic combinations of processes in different environments). Such a task is obviously difficult and existing classifications and ontologies are not sufficient to provide a description compatible with comparison, modelling and simulation of sophisticated processes. This is due to the fact that the available classifications have intrinsic limitations because they do not take into account the environment, elements of action shared between two functions, etc. For example, Gene Ontology (GO; Harris *et al.*, 2004) is becoming the standard in the field of classification of terms associated with biological processes. Although very useful, the description of biological processes in GO is limited to tagging molecules with standardized terms that are hierarchically classified. Likewise, the hierarchic enzyme commission (EC) nomenclature system (EC number) is used to classify enzyme reactions. It defines and categorizes catalysed reactions (Bairoch, 1993), but often the relation between EC number and enzyme is ambiguous. Indeed, a given enzymatic complex can sometimes be characterized not by a global EC number but by several EC numbers, because it contains several active sub-units that participate in different activities (e.g. the pyruvate dehydrogenase complex), as each of them catalyses a different reaction. Alternatively, the association of different polypeptides may be required to catalyse a given reaction and they will be all catalogued under a single EC number (i.e. citrate synthase). Moreover, a mono-domain enzyme has only one EC number, whereas it can have different actions in different contexts. Thus, there are different EC numbers for enzymes that have similar actions and the EC number cannot always represent properly the entire spectrum of enzymatic processes. It is worth noting that biological processes are many and heterogeneous and most of them can be decomposed into combinations of more simple processes, in which a process can be performed by part of a protein (domain), a single protein or a network. Thus, biological processes have to be described at different levels of abstraction. Moreover, the same process can be performed by different proteins depending on the context. As a consequence, it would be an advantage to describe processes in a generic form. For this, it is necessary to take into account the

*To whom correspondence should be addressed.

multi-functionality of a molecule, the context dependencies and the functional modularity. Thus, to describe biological processes in a relevant way there are some key requirements:

- Object and process separation: molecular objects and processes must be described separately.
- Modularity: the BioObject structure performing the processes should be described at different levels when possible.
- Multi-functionality: molecular objects can perform various functions.
- Context-dependency: any function depends on the context that constrains the biological processes.
- Elementariness: biological processes can be described by elementary actions.
- Genericity: the diversity of biological processes can be described as a limited set of generic process descriptions.

In the past, some attempts at describing biological processes provided a new viewpoint on process description but they did not fully integrate the associated constraints. For example, aMAZE (van Helden *et al.*, 2000, 2001) is a database for the representation of information on networks of cellular processes such as genetic regulation, biochemical pathways and signal transduction. It incorporates taxonomies for categorizing molecular entities and interactions at molecular, cellular and multi-cellular levels. It also makes a clear distinction between molecular entities and activities. Unfortunately, it does not decompose molecules into functional domains when relevant, despite the fact that protein domains now are considered as folding and functional units within proteins (Moore *et al.*, 2008).

Various computer science approaches have been adapted for the representation of biological systems [e.g. graph theory (Ravazs *et al.*, 2002), convex analysis (Schuster *et al.*, 1999), Boolean models (Bernot *et al.*, 2004), Petri nets (Matsuno *et al.*, 2000) and abstract machines (Calzone *et al.*, 2006)] but they do not fully integrate dynamics as well as molecular and biochemical details. Moreover, they always consider biomolecules and their functions as simplified interacting 'black balls' with automatic behaviours. These views are clearly not satisfactory for biologists given the level of detailed knowledge they can have on the functions of molecules. For example, BioCham (Calzone *et al.*, 2006) is a rule-based language for the representation of biomolecular systems, with a notation reminiscent of that of chemistry but the representation is a simplification of the biological knowledge. The Π-calculus (Milner, 1993), which is a formal computer language for describing concurrent computations, takes into account the sub-domain processes of proteins (Regev and Shapiro, 2002; Regev *et al.*, 2001). The Π-calculus modelling can formally represent detailed molecular and biomolecular information and allows their study with various simulations. A biological structure is represented by its potential behaviour: it manipulates molecules and protein domains as computational processes. This formalism is close to our expectation, but it does not separate the process from the object and the language is not convenient to integrate biological annotations. In most cases, either the processes are not generic (thus they need a huge variety of functional descriptions) or the biological knowledge is limited to the biological reactions, which are simplifications of the current detailed biological knowledge. In conclusion, biologists need a description scheme for the biological process that allows detailed and formalized annotations at different levels of abstraction compatible with the structural organization of the functional units (domains) of proteins. We propose a computational approach that makes use of the BioΨ concepts (Mazière *et al.*, 2004, 2007) allows the description of biological processes at four levels of abstraction from sub-molecular details to network modules. In addition, this approach is based on the principle of elementary bricks of actions and takes into account the biological context. BioΨ allows dissociating a biological entity, the actions that it performs and the context-dependency of such processes. Thus, biological processes can be defined in a generic form and are expressed in terms of a combination of a small set of elementary bricks of actions found in Nature. Herein, we propose a biological abstraction and a syntax based on BioΨ to provide the modelling community with a new powerful formalism that integrates biological complexity. This abstraction allows new computations with biological processes (comparison, prediction, etc.).

2 METHODS

Each molecule is considered as a biological object that can perform biological processes. Several essential aspects of biomolecular systems are identified that lead to a compositional model. Biomolecular systems are composed of a population of molecules that contains different molecular species with multiple copies of each type. Two types of molecules are distinguished: macromolecules [proteins, nucleic acids (NAs), etc.] and small molecules (metabolites, ions, etc.). Small molecules are treated as elementary objects, whereas macromolecules may be composed of several (functional) domains that may have one or more known interaction sites (Fig. 1).

We define four levels of structural description:

- A motif that constitutes an interaction site which can interact with another site or small molecule.
- A domain that contains a set of motifs.
- A molecular entity that contains a set of domains or a set of sub-units.
- A biochemical system that contains a set of molecular entities. Molecular entities can be organized in modules.

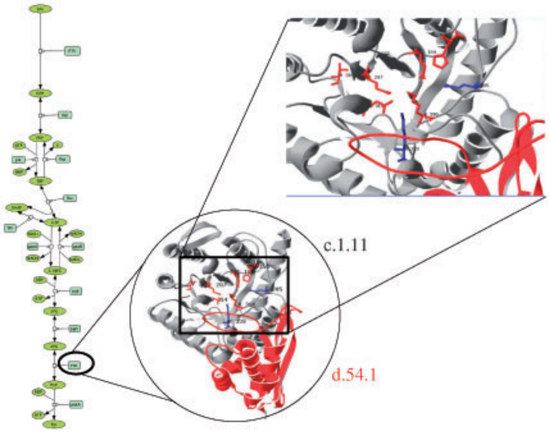

Fig. 1. Multi-scale description of the glycolysis pathway represented with Celldesigner (Funahashi *et al.*, 2008). As an example, enolase is an enzyme composed of two domains (c.1.11 and d.54.1) and the domain c.1.11 contains the substrate binding site (residues 155, 164, 287, 314, 339 and 390 in red) and the active site (residues 205 and 339 in blue).

In Figure 1, the biochemical system is the glycolysis pathway. It contains a set of molecular entities, one of which is the enzyme enolase. Enolase contains two domains; the first one, enolase_D1, is characterized by the SCOP fold c.1.11 and the second domain, enolase_D2, by the d.54.1 fold. Enolase_D1 contains the substrate binding site (in red) and the active site (in blue). A molecular entity made of two sub-units can be written in the same way. For example, the enzyme citrate synthase (EC 2.3.3.1) contains twice the same sub-unit that has a SCOP characteristic fold a.103.1 and the substrate binding and active sites are in the first sub-unit.

Similarly, most of the known biological processes can be decomposed into combinations of elementary processes and BioΨ has identified 97 basic element of actions (BEAs; Mazière *et al.*, 2004, 2007) at biochemical level for all species. BioΨ formalism allows: (i) the modelling of biological networks at different levels by taking into account the biological context based on elementary bricks of action; and (ii) the description of multi-level biological processes from sub-molecular details to networks. Thus, processes can be described at four levels that correspond to the structural views of molecules:

- BEAs refer to elementary actions at a chemical level that are involved in biological processes.
- Biological activities (BAs) represent the use of a combination of BEAs by functional domains to exert their activity at a sub-molecular level.
- Biological functionalities (BFs) represent the integration of the BAs of molecular entities.
- Biological roles (BRs) represent the combination of the BFs of different molecular entities within functional modules in the cell.

It is important to note that BEAs representation is not a chemical reaction equation, but a way to specify a type of elementary molecular modification. BEAs do not have any chemical reality on their own but identify elements of transformation from input to output molecules. They fall in four classes: bond modifications, transfers, chemical reorganizations and non-covalent actions. The complete list of the four BEA classes and the table of correspondence between EC number and BEA are at: http://www.sysdiag.cnrs.fr/publications/supplementary-materials/BioPsi_Manager/

3 SYNTAX

3.1 BioObjects

BioΨ manipulates formal objects that represent biochemical objects ranging from motifs, domains, molecular entities to modules of macromolecules. We define all these biological objects with the term of 'BioObjects'. In a formal way, biological processes at any level are represented by functions whose parameters are biochemical objects. Thus, the syntax of BioObjects can be expressed using the following grammar:

```
BioObjects = Smallmolecule | Macromolecule
Smallmolecule = Sugar | Lipid | Ion | AA
                | NA | Metabolite
Macromolecule = Protein | Complex | DNA
                | RNA
```

A BioObject can be either a small molecule or a macromolecule. Small molecules are considered as the simplest units among BioObjects because they cannot be decomposed into smaller ones; they can only be written with their atomic formulae composed of chemical atoms. They can be sugars, lipids, ions, amino acids (AAs), NA or metabolites. Macromolecules are the results of the combinations of small molecules as shown in Figure 1; they can be proteins, protein complexes, DNA or RNA molecules. The decomposition of a macromolecule can be written as follows:

```
Motif   = list_of_AA
Domain  = { seq:[int-int];
    fold:string;
        ml:list_of_Motif }
Protein = { an:string;
    dl:list_of_Domain }
Complex = { pl:list_of_Protein }
```

A motif is a list of AA; it is worth noting that this list can be empty if the site is unknown. A domain needs three fields to be defined: a fold represented by the SCOP number, the sequence represented by the UniProt number and the list of sites that can be empty if no interaction site is known. A molecular entity (protein) is formed by a domain or several domains, which are all part of the protein sequence. A molecular entity may be composed of several domains coming from separated polypeptidic chains. In such a case, it is treated as a multi-domain protein. A complex is formed by several associated proteins (which can be composed of several domains).

3.2 BioProcesses

A BioProcess can be a process at the chemical level (BEA), at sub-molecular level (BA), at molecular level (BF) or at cellular level (BR). The BioProcess syntax can be done with the following grammar:

```
BioProcesses = BEA | BA | BF | BR

BR =  BF | BF-1 | BR and BR
   | BR or BR | BR, BR
   | if (condition) do BR
BF = BA | BA-1 | BF and  BF
   | BF or BF | BF,BF
   | if (condition) do BF
   |parameters (list_of_paramaters): {BF}
BA = BEA | BEA-1 | BA and  BA
   | BA or BA | BA,BA
   | if (condition) do BA
   |parameters (list_of_paramaters): {BA}
BEA =  reactant -> reactant
reactant = BC
     | reactant + reactant
BC = atom | var | BC-BC
   | BC=BC | BC#BC | BC/BC
```

BioProcesses are defined in a generic form and are inductively expressed in terms of processes at the lower scale separated by logical operators. BioProcesses from BA to BR have a list of possible substrates. They may occur sequentially or concurrently. We abstract a sequence of processes with an operator ',' (comma sign), a conjunction with 'AND' and a disjunction with 'OR'. All the processes can be reversible (annotated by '-1') and they can be conditionally noted by the instruction 'if (condition) do' in which the condition is a Boolean value. The kinetic parameters can be associated with BF or BA depending on the knowledge (see example in Supplementary Section 3). A reactant can be a biological compound (BC) or several BCs separated by the operator '+'. The modifications of reactants is noted with '→'. A BC can be an atom, a variable or several BC connected with different kinds of bonds. Bonds can be simple '–', double '=', triple '#' or non-covalent '/'. For example, a bond modification with its reversible form can be written as follows:

Bea.CC.1(R) = R–COOH → R–H + CO
Bea.CC.1$(R)^{-1}$ = R–H + CO → R–COOH

R is a variable whereas C, O and H are chemical atoms. The first BEA describes the modification by which a CC bond becomes a CH bond with CO release; the second defines the inverse reaction. Supplementary Section 1 shows how to recast the phosphotransferase system process knowledge into its formal description.

3.3 Constraints and kinetics

The biological context of a given process can be transposed in a set of different constraints modulating its realization. These constraints mainly start to be applicable from the level of BA to BRs. The BEA level cannot integrate biological context since it is only relying on elementary pieces of function (more conceptual than real). Two kinds of constraints can be distinguished: constraints that condition the realization of processes and constraints that modulate the parameters of a process itself. The conditional constraints express eventual dependencies to specific localization or sequential process occurrence described using classical set of operators. The second type of constraints specifies the kinetic parameters associated with a given process. This can be implemented by the addition of available kinetic values to the set of parameters related to a molecular processes (from BA to BF; Mazière *et al.*, 2004).

4 COMPARISON WITH CURRENT EXISTING FORMALISMS AND ANNOTATION STANDARDS

Compared to BioΨ, GO and EC are only partially covering the four levels of abstraction ranging from biochemistry of reactions to the cellular level (see Supplementary Table 2). Enzyme classification are biochemical reaction oriented but does not use any level of abstraction. Moreover, when analysed, it covers without distinction the two first levels of abstraction of interest. On the opposite, GO uses the functional point of view and two levels of abstraction. On one hand, GO 'Biological processes relate to BioΨ' BRs, but on the other hand 'molecular function' fuses two levels of abstraction: BFs and BAs. Nothing is said about the biochemistry of the reaction. So, BioΨ is the unique annotation scheme providing domain level functional annotation close to the biologist point of view. As an example, Pfam in its last release (Finn *et al.*, 2008) intends to improve its synopsis of function annotation of protein domains. The authors raised the lack of standard format compatible to functional annotation at the domain level. Consequently, they remain using textual annotation.

Compared with current formalisms used to represent biological systems, Biocham and Π-calculus are the closer standard in terms of conceptual approaches. In Table 1, we compare BioΨ to these two formalisms. Only BioΨ and Π-calculus allow multi-level description of processes and generic enzymatic description. All the formalism are manipulating kinetics parameters. Since more ancient, Biocham and Π-calculus do benefit of existing tool to run processes. In the case of BioΨ, such tool is under development using the agent-based machine. On the opposite, BioΨ is the only one formalism to provide an elementary action-based description. Moreover, since the elementary action description came from a deep biological function analysis, the BioΨ process descriptions are closer to the biologist understanding and culture than Biocham and Π-calculus. If compared with current functional annotation standards used by the biologist community (GO, EC), BioΨ is better covering the sub-molecular description of processes found in nature (see Supplementary Table 2).

Table 1. Comparison of main representations of biological systems

	BioΨ	Biocham	Π-calculus
Multi-level	+	−	+
Generic enzymatic description	+	−	+
Kinetic parameters	+	+	+
Runnable with existing tools	under development	+	+
Elementariness	+	−	−
Proximity with biology	+	+	−

5 APPLICATION OF THE BIOΨ PRINCIPLES

5.1 Formal integration of processes

Formalization of biological processes eases their use for further calculations. All the elementary actions that form a BioProcess can be determined with the BioΨ formalism. In turn, all BioProcesses can be decomposed into lower scale processes composed of BEAs. Let $\Psi = \mathrm{BEA} \cup \mathrm{BA} \cup \mathrm{BF} \cup \mathrm{BR}$ be the set of all BioProcesses. If we note $\Psi_0 = \mathrm{BEA}$, $\Psi_1 = \mathrm{BA}$, $\Psi_2 = \mathrm{BF}$ and $\Psi_3 = \mathrm{BR}$, for all processes $p \in \Psi_i$ with $i \in \{1,2,3\}$ and $(q_j)_{1 \le j \le n} \in \Psi^n_{i-1}$, we have $f : \Psi^n_{i-1} \rightarrow \Psi_i$ such that $f(q_1,...q_n) = p$. The sequence of all BEAs of a process p is noted with the operator of restriction '$_|$' and is defined as

$$P_{|\mathrm{BEA}} = \{(b_i)_{1 \le i \le n} \in \mathrm{BEA} \text{ such that } f(b_i) = P\}$$

with $f : \mathrm{BEA}^n \rightarrow \Psi$.

For instance, using this expression, one can aggregate all the BEAs of a BF performed by a protein that is composed by two functional domains, each doing a different BA, as follows: $P \in \mathrm{BF}$, it exists $Q_1, Q_2 \in \mathrm{BA}$ and $f : \mathrm{BA} \times \mathrm{BA} \rightarrow \mathrm{BF}$ such that $P = f(Q_1, Q_2)$. So,

$$\begin{aligned} P_{|\mathrm{BEA}} &= Q_{1|\mathrm{BEA}} \cup Q_{2|\mathrm{BEA}} \\ &= (b_i)_{1 \le i \le k} \cup (b_i)_{k+1 \le i \le n} \\ &= (b_i)_{1 \le i \le n}. \end{aligned}$$

If the same BF is performed by a molecular entity made of only one functional domain, one can aggregate all the BEAs in the same way

$$\begin{aligned} P_{|\mathrm{BEA}} &= Q_{|\mathrm{BEA}} \\ &= (b_i)_{1 \le i \le n} \end{aligned}$$

Consequently, our formalism enables us to compare at molecular level the BEA composition of BF processes. This obviously can be extended to all levels of abstraction covered by BioΨ.

5.2 Comparison of biological processes: serine proteases

Thanks to the BioΨ principles biological processes can be compared in many different ways and at different levels of abstraction. Although it is affordable, it is very much dependent on the quality of the available functional knowledge concerning the studied molecular entities. Unfortunately, large-scale process comparison at all levels would need full process annotation at all these levels. However, process comparison at a given level can be performed. For instance, one can search for similarities between the process compositions of elementary actions.

Two processes with a comparable composition in elementary actions will be noted $=_{\mathrm{BEA}}$ if they have the same sequence of BEAs

$$p_1 =_{\mathrm{BEA}} p_2 \Leftrightarrow p_1|_{\mathrm{BEA}} = p_2|_{\mathrm{BEA}}$$

As an example, we will show how the BioProcesses of different serine proteases can be compared using BioΨ. Serine proteases are enzymes that cut specific peptide bonds in other proteins. This activity depends on a set of AA residues at the active site, one of that is always a serine. Serine proteases can be classified in different families. We consider here the trypsin-like serine proteases and the subtilisin-like serine proteases. Although their structural folds are different, their catalytic mechanisms are similar (Singer and Berg, 1997). The main player in the catalytic mechanism is a catalytic triad (His–Asp–Ser) and an oxyanion hole preserved in most serine proteases. These three key AAs play an essential role in the cleaving ability of serine proteases and their active site is shaped as a cleft to which the polypeptide substrate binds.

Chymotrypsin and trypsin are serine proteases of the trypsin-like family found in the digestive system of many vertebrates. They have very strong structural similarity and use the same cleavage mechanism; but each has its own selectivity as they cleave (Kassera and Laidler, 1969) different peptide bonds during protein digestion. Trypsin (EC 3.4.21.4) interacts with positively charged residues such as arginine (R) and lysine (K) on the substrate peptide to be cleaved. Chymotrypsin (EC 3.4.21.1) cleaves peptides at the carboxyl side of tyrosine (Y), tryptophan (W) and phenylalanine (F). Subtilisin (EC 3.4.21.62) is a serine protease of the subtilisin-like family secreted by the bacterium *Bacillus subtilis*. It hydrolyses proteins with a preference for a large uncharged residue in P1. It initiates the nucleophilic attack on the peptide (amide) bond through the serine residue at the active site. This information can be noted as follows:

$\mathrm{BF}_{\mathrm{Trypsin}} =$
$\mathrm{BA}_{\mathrm{serprot}}(\text{[RK]-x, } \mathrm{H_2O}\text{, } \mathrm{Ser}^{195}\text{, } \mathrm{His}^{57}\text{, } \mathrm{Asp}^{102})$
$\mathrm{BF}_{\mathrm{ChymoTrypsin}} =$
$\mathrm{BA}_{\mathrm{serprot}}(\text{[YWF]-x, } \mathrm{H_2O}\text{, } \mathrm{Ser}^{195}\text{, } \mathrm{His}^{57}\text{, } \mathrm{Asp}^{102})$
$\mathrm{BF}_{\mathrm{Subtilisin}} =$
$\mathrm{BA}_{\mathrm{serprot}}(\text{[YWTCSNQFMLIVGAP]-x, } \mathrm{H_2O}\text{, } \mathrm{Ser}^{221}\text{, } \mathrm{His}^{64}\text{, } \mathrm{Asp}^{32})$

$\mathrm{BA}_{\mathrm{serprot}}(s_1, s_2, aa_1, aa_2, aa_3) =$
$\mathrm{Ba.CO.1}^{-1}(aa_1, s_1)$,
$\mathrm{Ba.lab.2}^{-1}(aa_2)$ and $\mathrm{Ba.CN.2}(s_1)$,
$\mathrm{Ba.lab.1}(aa_3)$ and $\mathrm{Ba.lab.1}(s_2)$ and $\mathrm{Ba.CO.1}^{-1}(s_2)$,
$\mathrm{Ba.CO.1}(aa_1)$
with
Ba:CO.1: $\mathrm{C–O–R} \rightarrow \mathrm{C}^0 + \mathrm{R–O}^0$
Ba:CN.2: $\mathrm{C–N(R')–R''} \rightarrow \mathrm{R'–N^0–R''} + \mathrm{C}^0$
Ba:lab.1: $\mathrm{H–OH} \rightarrow \mathrm{H}^0 + \mathrm{HO}^0$
Ba:lab.2: $\mathrm{R–H} \rightarrow \mathrm{R}^0 + \mathrm{H}^0$

$$\begin{array}{lcl}
\mathrm{Subtilisin}_{|\mathrm{site}} & = & \{\mathrm{Asp}^{32};\ \mathrm{His}^{64};\ \mathrm{Ser}^{221}\} \\
\mathrm{Subtilisin}_{|\mathrm{fold}} & = & \mathrm{c.41.1} \\
\mathrm{Trypsin}_{|\mathrm{site}} & = & \mathrm{Chymotrypsin}_{|\mathrm{site}} \\
 & = & \{\mathrm{His}^{57};\ \mathrm{Asp}^{102};\ \mathrm{Ser}^{195}\} \\
\mathrm{Trypsin}_{|\mathrm{fold}} & = & \mathrm{Chymotrypsin}_{|\mathrm{fold}} \\
 & = & \mathrm{b.47.1}
\end{array}$$

Trypsin/chymotrypsin and subtilisin use the same catalytic triad mechanism at their active site despite having different structures (Martin *et al.*, 1998). This is the classic example used to illustrate convergent evolution, since the same mechanism evolved twice independently during evolution: two molecules acquired the same function (analogous) despite having evolved from different genes. The folds c.41.1 and b.47.1 have a common biological activity.

5.3 Integration of biological complexity: glycolysis

Most of the proteins, which exhibit significant structural similarity in terms of folding, are homologous and perform similar or identical BA. Proteins and enzymes are involved in biological networks in which they carry out their potential activity depending on the context. If one describes for each enzyme of a network its BFs, BAs and BEAs composition, then it is possible to compare protein networks based on their BEAs composition (Mazière *et al.*, 2007). For instance, the glycolysis pathway, which is composed of 13 enzymes, can be expressed in such a way (see Supplementary Materials). In this way, biological processes that participate in a given biological network can be expressed in a more formal way compared to the classical textual description found in the current databases (Swissprot, Pfam, etc.). Unfortunately, knowledge on the parameters of biological processes is not always complete or available. Nevertheless, BioΨ multi-level description allows to cope also with incomplete knowledge since it is possible to aggregate information at a hierarchical lower level to the next upper level. This feature can be used when details about an intermediate level are missing. For example, since the activity of each sub-unit of the homotetramer phosphofructokinase is not known (see Supplementary Materials), the global activity is annotated for the whole enzyme.

6 DISCUSSION

To compute with biological functions, we need a formalism that on one hand, can integrate part of the complexity of the existing knowledge, and on the other hand, can allow us to make calculations with biological processes.

The biological knowledge on functions is wide and very heterogeneous but very little is available in databases about such biological processes despite the fact that a considerable wealth of annotations exist on the biological objects (proteins, NAs, small molecules, etc.) that support such processes. In others words, we know more about the players of a process than about the process itself. In addition descriptions of functions and BioObjects are often confusing. BioΨ clearly separates process description and BioObject description and focuses on the different levels, going from the chemical, sub-molecular and molecular entities to the network level. All other existing function annotations do not or only partially address this view. BioΨ is based on the integration of all these levels and it could extend to additional levels if required. We described BEA at the chemical level, which can be considered as the elementary bricks that compose all functions in Nature. BEAs fall into four classes and their relationships with EC are available at BioΨDB (http://www.sysdiag.cnrs.fr/publications/supplementary-materials/BioPsi_Manager/). Obviously, the description of elementary actions provides a generic view of a biological process. It can be used as an element of comparison. To integrate available knowledge on functions, we can complement the description of a biological process with parameters such as substrate specificity, kinetics, conditionality of the process execution based on contextual information (pH, temperature, etc.). These parameters can be global (i.e. temperature) or local to a given level (kinetics, substrate specificity, etc.). The BioΨ paradigm to describe such biological processes is expressed herein in a formal way in order to allow further calculations. This formalism tolerates missing annotations at intermediate levels as, in this case, it can aggregate lower level information into the first upper level where information is available. It is worth noting that the ability of BioΨ to provide generic descriptions (the composition of elementary actions in a given sequence) reduces drastically the combinatorial aspect observed in classical descriptions of functions. This is a great advantage especially if large-scale functional annotations are envisaged. For instance all kinase processes, which are currently described in very heterogeneous ways and have different EC numbers, are based on the same (or similar) sequences of BEAs. Only the specific parameters at the various levels will be different. The BioΨ approach provides a new tool for systematic mapping of structure–function relationships. Most of the proteins with the same topology are homologous and have similar functions. A fold f is in relation with a BEA b, if there is a BioObject o such that $o|_{\text{fold}} = f$ and $b \in o|_{\text{BEA}}$. This link allows a formal construction of a structure–function map to propose and test hypothesis on the possible functions of a given structure or of a possible folding in a given protein with a given function. In the future, a wide map of specific links at different levels of details between the BioObject structural elements and its elementary actions might open the way for new methods of structure–function predictions. Moreover, BioΨ formal description allows the comparison of systematic processes without going through the filter of AA (or NA) sequence comparison. Hence, one can expect less bias in the comparison since it will overcome the bias introduced by domain shuffling in sequences. We also would like to stress that the notion of elementary action is very much coherent with the genetic mechanisms of evolution in Nature. Since the sequences of macromolecules are made of reused blocks that have been shuffled, mutated, truncated, etc. during evolution, then it is not surprising to observe limited sets of typical blocks (i.e. folds) in their 3D structures. Similarly, one could expect that these sequence-structure blocks could support limited sets of elementary actions. Thus, diversity in Nature is mainly the result of the combinatorial use of limited blocks rather than the outcome of the diversity of the composing blocks. Other uses of BioΨ are under development in the context of modelling of complex systems. Here, BioΨ descriptions are used as hubs for biological detailed knowledge, modelling and simulation tools. For instance, we can extract sets of data from BioΨ descriptions and rewrite them in the various formats used in modelling (system biology markup language, elementary flux modes, ordinary differential equation etc.). BioΨ carries strong advantages to run agent-based simulations. Multi-level will allow various angles of view of simulation with a unique model description. Elementariness may prevent from combinatorial explosion in function description. Finally, sub-molecular function description open the doors for further non-deterministic simulation, where neoformed pieces of molecules can carry their own new function. A new tool is currently under development in this field.

ACKNOWLEDGEMENTS

We thank Patrick Amar for fruitful discussions.

Funding: BaSysBio project.

Conflict of Interest: none declared.

REFERENCES

Bairoch,A. (1993) The ENZYME data bank. *Nucleic Acids Res.*, **21**, 3155–3156.

Bernot,G. *et al*. (2004) Application of formal methods to biological regulatory networks: extending Thomas' asynchronous logical approach with temporal logic. *J. Theor. Biol.*, **229**, 339–347.

Calzone,L. *et al*. (2006) BIOCHAM: an environment for modeling biological systems and formalizing experimental knowledge. *Bioinformatics*, **22**, 1805–1807.

Finn,R. *et al*. (2008) The Pfam protein families database. *Nucleic Acids Res*., **36**, D281–D288.

Funahashi,A. *et al*. (2008) CellDesigner 3.5: a versatile modeling tool for biochemical networks. *Proc. IEEE*, **96**, 1254–1265.

Harris,M.A. *et al*. (2004)The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, 258–261.

Karp,P. (2000) An ontology for biological function based on molecular interactions. *Bioinformatics*, **16**, 269–285.

Kassera,H. and Laidler,K. (1969) Mechanisms of action of trypsin and chymotrypsin. *Can. J. Chem.*, **47**, 4031.

Martin,A. *et al*. (1998). Protein folds and functions. *Structure*, **6**, 875–884.

Matsuno,H. *et al*. (2000) Hybrid petri net representation of gene regulatory network. *Pac. Symp. Biocomput.*, **5**, 341–352.

Mazière,P. *et al*. (2004) A biological processes description scheme based on elementary bricks of action. *J. Mol. Biol.*, **339**, 77–88.

Mazière,P. *et al*. (2007) Formal description of TCA cycle based on elementary bricks of action. *J. Biosci.*, **32**, 145–155.

Milner,R. (1993) The polyadic π-calculus: a tutorial. In Bauer,F.L. *et al*. (eds), *Logic and Algebra of Specification*. Springer-Verlag, pp. 203–246.

Moore,A. *et al*. (2008) Arrangements in the modular evolution of proteins. *Trends Biochem. Sci.*, **33**, 444–451.

Ravazs,E. *et al*. (2002) Hierarchical organization of modularity in metabolic networks. *Science*, **297**, 1551–1555.

Regev,A. and Shapiro,E. (2002) Cellular abstractions: cells as computation. *Nature*, **419**, 343.

Regev,A. *et al*. (2001) Representation and simulation of biochemical processes using the pi-calculus process algebra. In *Proceedings of the Pacific Symposium of Biocomputing*. Vol. 6. World Scientific Press, Singapore, pp. 459–470.

Rison,S. *et al*. (2000) Comparison of functional annotation schemes for genomes. *Funct. Integr. Genomics*, **1**, 56–69.

Schuster,S. *et al*. (1999) Detection of elementary modes in biochemical networks : a promising tool for pathway analysis and metabolic engineering. *Trends Biotechnol.*, **17**, 53–60.

Singer,M. and Berg,P. (1997) *Exploring genetic mechanisms*. University Science Books, USA.

van Helden,J. *et al*. (2000) Representing and analysing molecular and cellular function using the computer. *Biol. Chem.*, **381**, 921–935.

van Helden,J. *et al*. (2001) From molecular activities and processes to biological function. *Brief. Bioinform.*, **2**, 81–93.