APPROACHES TO CHEMINFORMATICS AND BIOMATERIALS DESIGN BY NEURAL NETWORKS FOR STRUCTURES: FIRST APPLICATIONS TO SMALL MOLECULES AND POLYMERS

A. Micheli, A. Starita, C. Duce, R. Solaro, M. R. Tiné
University of Pisa, Italy

ABSTRACT

We present an overview of a new approach to cheminformatics based on neural networks for structures. In particular, we show the relevance of this methodology in the wider framework of new method development for biomaterial design and medicinal chemistry purposes. Current advancements, which include application to the prediction of properties for both small molecules and macromolecules, show the generality and flexibility of the proposed approach.

Keywords: recursive neural networks, cheminformatics, biomaterials, QSPR/QSAR.

INTRODUCTION

The aim of this paper is to put in evidence the evolution and potentiality of emergent computational intelligence approaches exploiting recursive neural networks for the applications to medical cheminformatics. After introducing the main motivations to collocate the current short-term studies into a wider frame of long-term medical research results, an overview of the problem domain and of the methodology will be provided. Finally, a summary of the qualitative results of the current research, applied to different tasks, showing the potentiality and generality of the new approach will conclude the paper.

Biomaterials and Cheminformatics. A major challenge confronting pharmaceutical chemists is the rationale design of drug molecules to optimize pharmacological interactions with their therapeutic targets and to enable them to circumvent biological barriers (e.g., intestinal mucosa, liver, blood-brain barrier) that separate the site of drug administration from the site of drug action [1]. The inability to circumvent such barriers often prevents leading drug candidates from being clinically developed. Therefore, scientists in the pharmaceutical industry employ *in situ* (e.g., perfused organs) and *in vitro* (e.g., tissue and cell cultures) tests in order to optimize both the pharmaceutical properties of drug candidates and the carrier systems. In fact, the ways in which drugs are administered also play a fundamental role in determining the therapeutic efficacy of a drug. Conventional dosage forms are not able to control either the rate of drug delivery or the target area of drug administration and provide an immediate or rapid drug release. This necessitates frequent administration in order to maintain a therapeutic level. As a result, drug concentrations in the blood and tissues fluctuate widely. The concentration of drug is initially high, that can cause toxic and/or side effects, then quickly falls down below the minimum therapeutic level. The duration of therapeutic efficacy is dependent upon the frequency of administration, drug half-life, and release rate from the dosage form. In contrast, controlled release systems are not only able to maintain therapeutic levels of drug with narrow fluctuations but they also make it possible to reduce the frequency of drug administration [2].

One way of modifying the biodistribution of drugs is to entrap them in submicroscopic drug carriers as nanospheres or nanocapsules [3]. Nanoparticles are solid colloidal particles made of synthetic or natural polymers with diameter ranging from about 10 to 1,000 nm, in which biologically active molecules can be entrapped, dissolved, or encapsulated, and/or to which the active principle is adsorbed or attached. Nanospheres have a matrix type structure, whereas nanocapsules have a polymeric outer shell and an inner liquid core [4]. Nanoparticles can be designed for different kind of administration routes: intravenous, intramuscular, subcutaneous, oral, nasal, ocular, transdermal. The size and surface characteristics of nanoparticles, in terms of charge, hydrophilic-hydrophobic balance, and presence of reactive groups will dramatically affect their body distribution and targeting attitude [5]. The sub-micron size allows nanoparticles to penetrate deep into tissues through fine capillaries, to cross the fenestration present in the epithelial lining, and to be taken up for drug delivery. Moreover, drug-loaded nanoparticles exposing targeting moieties could effectively and selectively drive the active principle at the desired site of action with substantial abatement of side effects.

It must be stressed that a controlled release system comprises both the drug and the material in which the drug is loaded. Therefore, the selection of the drug and the polymer along with desired properties is a prime factor in designing a controlled release system. In particular, the pursuit of an adequate compromise of bulk and surface properties represents an important issue to be addressed. Among the general characteristics that drug delivery systems should present, it is possible to mention the ability to incorporate the drug without damaging it, tuneable release kinetics, long *in vivo* stability, lack of toxicity, carcinogenicity, and immunogenicity, potential of targeting specific organs and tissues. They should be also free of contaminants and leachables. It has also to be considered that the biological performance of a material depends on the host response to the biomaterial as well as to the material response to the living system. Biocompatibility tests identify any reactions that would lead to material failure or would result in disease. These effects include irritation, inflammation, pyrogenicity, interaction with blood, carcinogenicity, mutagenicity, systemic toxicity, sensitization, and reaction to foreign bodies.

On a parallel research line, over the last few decades, there has been a marked development in science and technology of materials designed for biomedical implants. Many clinical justifications exist for the employment of implants. They are needed for the removal of congenital defects and for the replacement of tissues that have been either damaged or destroyed by pathological processes. Prosthetics and biomedical devices are fabricated from biomaterials and surgically inserted into the living body. They can serve as permanent or temporary replacement of body parts [6]. In the first case, medical devices may replace a damaged part of anatomy, e.g., total joint replacement; simulate a missing part, e.g., mammary prosthesis; correct a deformity, e.g., spinal plates; aid in tissue healing, e.g., burn dressings; rectify the mode of operation of a diseased organ, e.g., cardiac pacemakers; or aid in diagnosis, e.g., insulin electrodes. In the second case, the implants are intended to function in the body for some time in order to restore a tissue or an organ.

In the early 1930s, the only biocompatible materials were wood, glass, and metals. These were used mostly in surgical instruments, paracorporeal devices, and disposable products. The advent of synthetic polymers and biocompatible metals in the latter part of the twentieth century has changed the entire character of health care. Polymers, metals, and ceramics originally designed for commercial applications have been adapted for prostheses, opening the way for implantable pacemakers, vascular grafts, catheters, and a variety of other orthopaedic devices. In recent years, polymeric biomaterials have gained increased importance through object-oriented synthesis, blends, and modifications that produce tailor-made characteristics for the areas where these materials are to be used.

Tissue engineering is a relatively new and emerging interdisciplinary field that applies the knowledge of bioengineering, life sciences, and clinical sciences for trying to solve the critical medical problems of tissue loss and organ failure [7]. It involves applying engineering principles of transport and reaction phenomena as well as methods of analysis aimed at understanding the complex biological processes that occur in tissue development and repair. Tissue engineering exploits living cells into a variety of ways to restore, maintain, and enhance tissues and organs [8]. Indeed, engineered tissues could reduce the need for organ replacement as well as accelerate the development of new drugs thereby eliminating the need for organ transplants. To engineer living tissues *in vitro*, cultured cells are coaxed to grow on bioactive degradable scaffolds that provide the physical and chemical cues to guide their differentiation and assembly into three-dimensional structures. Materials used for tissue engineering applications must be designed to stimulate specific cell response at molecular level. They should elicit specific interactions with cell integrins and thereby direct cell proliferation, differentiation, and extracellular matrix production and organization. A careful control of the topochemical microstructure of the material surface is strongly required to accomplish this goal. Indeed, the selection of biomaterials constitutes a key point for the success of tissue engineering practice. Moreover, these products must retain their functions effectively and safely over the desired period of time, without irritation of the surrounding tissue by either mechanical action or possible degradation products. This is ensured only when the biomaterial is biocompatible. Several overlapping processes determine biocompatibility. Not only the material mechanical and chemical-physical characteristics but also the special place of application, the individual reaction of the complement system, and the cellular immune system as well as the physical condition of the patient influence the material tolerance.

In tissue engineering, the chemical and physical characteristics of the biomaterial surface, which are responsible for the biological reactions at the interface are certainly of great importance. Influencing factors are the surface chemical structure, hydrophilicity, morphology, and the topography [9]. Surface characteristics can considerably differ from polymer bulk characteristics. Due to the minimization of the surface energy, the non-polar groups move to the phase boundary with air [10]; migration of low molecular components leads to differences between the surface and the bulk [11]. At the phase boundary between the biomaterial and the aqueous surroundings of the tissue, the situation is very different from that between the biomaterial and air. Thus, the surface characteristics can considerably change after the biomaterial is taken from an air medium into an aqueous system.

When the implant is exposed to the biological system, the following reactions are observed: (i) within the first few seconds, proteins are deposited from the surrounding body liquids. The structure of the adsorbed proteins is dependent on the surface characteristics of the implanted material. Additionally, the adsorbed proteins are subject to conformational changes as well as exchange processes with other proteins [12]. (ii) The tissue in contact with the implant reacts with dynamic processes that are comparable to body reactions in cases of injuries or infections. Due to mechanical and chemical stimuli, the implant can lead to a lasting stimulus of inflammation processes. As a result of being accepted by the organism, a biocompatible implant should thus be surrounded with a thin tissue layer, which is free of inflammation cells [13]. (iii) During the course of the contact between the biomaterial and the body, the aggressive body medium will cause degradation processes. Hydrolytic and oxidative processes can lead to the loss of mechanical stability and to the release of degradation products [14].

(iv) Because of the transport of soluble degradation products through the lymph and vessel systems, a reaction of the whole body towards the implant cannot be excluded. Infection of the biomaterial with bacteria has to be considered as an additional obstacle [15].

The description of the factors that together determine the biocompatibility of an implant shows the diversity of the processes. Moreover, while the term biocompatibility refers to the tolerance of biomaterials with liquid or solid body elements, the term hemocompatibility defines the tolerance of biomaterials with blood. Due to the enormous demand for implants and medical-technical goods for the cardiovascular area, blood tolerance is of great importance. From a clinical point of view, a biomaterial can be considered as blood compatible when its interaction with blood does not provoke either any damage of blood cells or any change in the structure of plasma proteins [9]. As a consequence of the non-specific protein adsorption and adhesion of blood cells, the contact of any biomaterial with blood often leads to different degrees of clot formation [16]. The competitive adsorption behaviour of proteins at the biomaterial surface determines the pathway and extent of intrinsic coagulation and adhesion of platelets. Predictions about the interactions between the biomaterial surface and the adsorbed proteins can only be formulated by having an exact knowledge of the structure of the biomaterial surface and of the conformation of the adsorbed proteins [17].

By examining the polymeric materials that are currently used in clinical application, it can be seen that while their mechanical properties satisfy requirements, their total compatibility with blood has still not been achieved. Therefore, commercial polymers, which are used as short-term implant materials, show thrombogenic properties and require the introduction of anticoagulants [18]. As a result of the complex interactions between the implant and the tissue, the expectation of unsatisfactory implant biocompatibility and/or hemocompatibility is high. Until now, it has not been possible to quantitatively understand these processes, and to relate them to the chemical structure of the biomaterial. Thus the design and the development of suitable biomaterials for successful use of implants is difficult. At the same time, in the field of drug design, the development of combinatorial chemistry has allowed the synthesis of a lot of compounds of medical interest, but the application of experimental test is cumbersome, expensive, and time consuming. Consequently, the development of predictive methods to evaluate candidates for specific applications has gained urgency both in drug delivery and in tissue engineering fields.

In time, significant efforts have been spent on the development of Quantitative Structure-Activity/ Property Relationship (QSAR/QSPR) techniques in order to predict the physical, chemical, biological, biomedical, and technological properties of molecules. The aim of a QSAR/QSPR approach is to find an appropriate function, *F*, which given a proper representation of a molecule predicts its biological activity or a selected property, as in the following:

$$\textit{Property} = F(\textit{Structure})$$

The function *F* can be described as the sequential solution of two main problems (see e.g. [19-21]): (i) the feature representation problem, i.e., how to encode molecules through the extraction and selection of structural features; and (ii) the mapping problem, usually faced by linear or non-linear regression tools (i.e., a *mapping function*). In more detail, the feature representation process requires the solution of two subtasks: the first one for the explicit representation of the significant structural information carried by molecules, and the second one for the encoding of this structural information into a numerical representation (by an *encoding function*).

Traditional QSAR/QSPR approaches, employing standard regression methods (from linear regression to standard neural network) take as input fixed-size numerical vectors. As a consequence, all molecules must be reduced to vectors of the same dimension by using a suitable group of molecular descriptors. The molecule can be represented by using different encoding approaches, such as, the selection of physical-chemical, geometrical, and electronic properties; the calculation of topological or connectivity indices; the occurrence of each group in the molecular structure. The need for molecular descriptors limits the type of modelled molecules (for example there are no descriptors for inorganic metal complexes) and determines the applicability of the method. As a matter of fact, the number and types of numerical descriptors used to represent chemical compounds are strictly dependent on the (*target*) property under study; for this reason, the models are not target-invariant. In particular, an expert has to start again from scratch the process of choosing suitable descriptors whenever a different property is investigated.

The central point of our analysis stems from the fact that molecules are not simply fixed-size vectors of numbers but they are more naturally described via a varying size structured representation. Beside specific success and the capability to abstract from expensive experimental test, the direct treatment of molecules in their natural form of structured objects is still an open issue of the current QSPR/QSAR approaches. To overcome this problem we introduce the use of the Recursive Neural Network, a model able to predict the desired properties directly from a structured representation of molecules.

NEURAL NETWORK FOR STRUCTURES

Before entering in the description of the results, let us introduce the basics of the Recursive Neural Networks (RNN) model we used to tackle the learning in structured domain for chemical data. A recurrent neural network distinguishes itself from a feedforward network (Multi-layer Perceptron) by having feedback loop connections in its topology, i.e. a weighted version of the output is fed back into the input. The presence of feedback provides the neural model with dynamic properties, by the use of contextual internal states (*memory*). The first exploitation of this concept concerned recurrent neural networks able to deal with sequences by a dynamic learned internal memory.

The recursive neural network is the generalization of the recurrent model to deal with more complex structures, e.g. labelled trees and labelled DPAGs (Directed Positional Acyclic Graphs) [22]. In such structures, for each vertex (or *node*) a total order is defined on the edges leaving from it and a position is assigned to each edge. We assume a bounded out-degree and that each DPAG possesses a super-source, i.e. a vertex *s* such that every vertex in the graph can be reached by a directed path starting from *s. Labels* are tuples of variables attached to vertexes. Let $\Re^n$ denote the label space. In particular, *k-ary trees* (*trees* in the following) are rooted positional trees with finite out-degree *k,* i.e. *k* is the maximum number of children for each node. The supersource is the root of the tree.

In the framework of the QSPR/QSAR analysis, and according to the RNN architecture, the processing of a RNN can be presented as the sequential application of two functions, an *encoding function* and a *mapping function.* Let us consider a realization of the two functions by a recursive neural network with *m* hidden neurons, i.e. a fully connected recursive neural network with one hidden layer. The *encoding* of an input structure, e.g. a tree T, is made by the hidden units computing for each vertex of T a numerical code ($\boldsymbol{x}$ in $\Re^m$) using information both of the vertex label ($\boldsymbol{l}$ in $\Re^n$) and, recursively, of the code, denoted as $\boldsymbol{x}^{(j)}$ in $\Re^m$, of the sub-trees descending from the current vertex. The *encoding function*, i.e. the output $\boldsymbol{x}$ of the hidden units for a vertex *v* (the code of *v*), is computed as:

$$\boldsymbol{x} = \Phi(\boldsymbol{W}\,\boldsymbol{l} + \sum_{j=1}^{k} \hat{\boldsymbol{W}}^{j} \boldsymbol{x}^{(j)}) \qquad (1)$$

where Φ is a set of *m* sigmoid functions, $\boldsymbol{W}$ in $\Re^{m\times n}$ is the weight (free-parameters) matrix associated with the label space, and $\hat{\boldsymbol{W}}^{j}$ in $\Re^{m\times m}$ is the weight (free-parameters) matrix associated with the *j*-th sub-tree space. The *bias* is included in the label $\boldsymbol{l}$. Through Equation 1 the encoding function is recursively computed for all the vertexes of the input structure and a code for the whole structure is returned at the root.

An instance of this model with *m=1*, i.e. a single *recursive neuron* unit, is graphically shown in Figure 1: the current information is expressed by the label field (l_1,..., l_n) of the vertex. Note that the vector $\boldsymbol{x}^{(j)}$ can be considered an extension of the inputs to the standard neuron that store the information from previous outputs of the model. The extended inputs represent "context" information about the subgraphs of the current processed input vertex. The weights $\hat{\boldsymbol{W}}^{j}$, *j=1,...,k* are specific of the *recursive* neuron (with respect to the standard one) and they are the free parameters associated to the stored sub-tree codes.

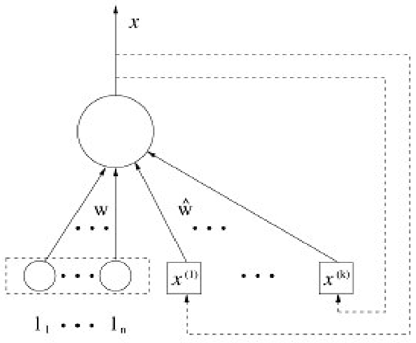

Figure 1: A single recursive unit.

Different architectures of the neural network that realize the encoding function can be considered. In particular, in the following we used a constructive approach, a Recursive Cascade Correlation method [21,22], which adds the hidden recursive neurons during the training of the model. Since this method automatically determines the number *m* of hidden units, it has been found particularly useful in the applications when no information is given on the complexity of the problem. In order to realize the output *mapping function* for the regression model, we use a single linear output neuron:

$$y = g(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x} + \theta \qquad (2)$$

where $\boldsymbol{A}$ is a weight (free-parameters) matrix in $\Re^m$ and θ is the output threshold.

The encoding process of the RNN is graphically represented in Figure 2 for two input structures representing 1-methoxypropane and 2-methyl-2propanol, respectively.

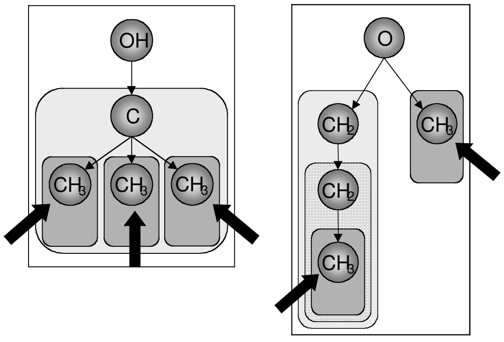

Figure 2: Unfolding the encoding process through structures. Each box includes the sub-trees progressively encoded by the recursive neural network.

Note that the encoding process mimics the morphology of each compound. As shown in Figure 2, the encoding is a bottom-up process starting at the leaves (black arrows in Figure 2). This corresponds to a visit (*traversal*) of the input tree according to an inverse topological order.

The model described in Equation 1 is applied for each step of the traversal. For each vertex, the model uses the label of the vertex and, recursively, the encoding value of the subtrees descending from the current vertex (depicted by the boxes in Figure 2). At the root, this process computes a code of the whole molecular structure. The code is then mapped to the output property value by the output function $y=g(\boldsymbol{x})$.

We can summarize the characteristics of the RNN approach in the context of QSPR/QSAR application by the following main points:

- RNNs take directly a structured representation of the molecules as input;
- The recursive models can learn (tuning the free-parameters) how to encode the structured representation of the molecules according to the given QSPR/QSAR task;
- Through the *encoding* and *mapping* function the RNN models a direct and adaptive relationship between molecular structures and target properties;
- Hence, RNNs discover by learning the specific structural descriptors (numerical code) for the QSPR/QSAR task at hand. As a result, no *a priori* definition and/or selection of properties by an expert are needed.

APPLICATIONS OVERVIEW

The first successful applications of this model were achieved predicting the boiling points of linear and branched alkanes and the pharmacological activity of a series of substituted benzodiazepine [19-21,23]. More recently, further advancements have been done to deal with a widest set of molecular structures and to address different chemical tasks [24,25].

In particular, we tested the RNN-based method for small molecules by applying it to the prediction of standard free energy, $\Delta_{solv}G°$, of solvation in water of a set of almost 300 linear mono- and poly-functional organic compounds. Afterwards, according to the goals described in the Introduction, we extended the method to macromolecules by investigating the glass transition temperature, Tg, of a set of acyclic hydrocarbon-chain polymers. The wide class of different chemical data and QSPR/QSAR problems faced by these applications provide the support to show the generality of the approach.

Concerning the first task, we considered the prediction of the solvation free energy of small organic molecules that is a parameter of considerable interest in drug design [26], in the analysis of protein folding and binding [27], and in the development of force fields by computer simulation [28-30]. In fact, the standard solvation free energy of a solute A in water and in a selected immiscible organic solvent, can be related to the logarithm of partition coefficient, logP, of A between the two liquid media. The partition coefficient is of critical importance for solvent extraction, in environmental and pharmaceutical applications. As a matter of fact, the transport of a molecule from the aqueous medium of the extracellular region to the cytoplasm is often ruled by passive diffusion across the lipid bilayer of the cell membrane and it is closely connected to its lipophilicity. For a given solute, this property is usually expressed by logP between water and a suitable apolar organic solvent [31-35]. In other words, the bioactivity of a molecule can be related to its logP. Furthermore, solvation properties were selected as the target property because of the availability of a large dataset of reliable literature data. Indeed, a homogeneous and critically reviewed data base is needed in order to reliably assess which performances may come from the application of the proposed model to a given problem.

In our approach, the representation of a molecule is directly derived from its molecular structure alone. For this purpose we describe the molecule as a 2-D graph that can be easily obtained from the structural formula. The investigated molecular structures were represented in terms of labelled rooted trees (k-ary trees), which are the subclasses of DPAGs covering the investigated structures. To this aim we studied an appropriate set of rules in order build a unique chemical tree for each molecule. Since labelled structures are high abstract and graphical tools we could describe a molecule at different levels of detail, such as atom bonds and/or chemical groups. In particular, each compound was divided into defined atomic groups. Each group corresponds to a vertex of the tree and each bond between them corresponds to an edge. We chose the smallest number of atomic groups able to build the greatest number of molecules in a reasonably compact form. The labels of vertex are categorical attribute distinguishing the symbols of the atomic groups. We decided to place the root in the functional group characteristic of the class to which the molecule pertains, all the other groups descending from it are considered branches.

The results obtained by using the proposed model were satisfactory. Various splittings of the data were used for training and validation purposes. We found that the data used for training the system were reproduced within the experimental error; the data of test sets were predicted with a mean absolute error and standard deviation lower than those reported in the literature for standard QSPR methods (for details see refs. 24 and 25).

Over the years, theoretical studies performed on the solvation process have allowed the identification of electronic, superficial, structural, and chemical reactivity characteristics that concur in determining the solvation free energy. By having in mind these information, we investigated the RNN learning process by principal component analysis (PCA) of the internal representation of molecules (i.e. the output of the encoding function) built by the neural network.

The PCA analysis has shown that the RNN is able to cluster the molecules not only by considering the chemical similarity of the molecular trees, but also by abstracting chemical information from the relationships between structures and targets learned by the model. For instance, the solvent accessibility of polar groups and their ability to act as hydrogen bond donor and/or acceptor are responsible of the distribution of polar molecules in the representation space developed by the RNN. However, it has been observed that the chemical knowledge abstracted by the model cannot be trivially decoupled into single effects. On the contrary, the model combines the structural and chemical features of the molecules by developing a sort of "smooth rule", reflected by the spread of the points in the clusters, globally accounting for the complexity of the stereo-electronic properties of molecules.

The success obtained with small molecules encouraged us to extend the approach to the prediction of polymer properties. In the present research, the glass transition temperature of a set of acyclic polymers including polyacrylates, polymethacrylates, polyacrylamides, polymethacrylamides, and some α- and β-substituted polyacrylics and polymethacrylics was investigated. Acrylic and methacrylic polymers were chosen because of the availability of a large number of experimental data, which allows for testing the potential of our RNN model with macromolecules. On the other hand, it is well known that the glass-rubber transition is of considerable technological significance. In fact, the Tg determines the utilization limits of rubbers and thermoplastic materials. For instance, the Tg of materials designed to replace soft and hard tissue must be lower than and well above body temperature, respectively.

In order to predict polymers Tg property, standard regression methods are reported in literature [36-43]. These methods use molecular descriptors for the representation of the molecular structure. The limitations associated with standard QSAR/QSPR methods already evidenced in the treatment of small molecules are exacerbated in the case of polymers. Molecular descriptors are indeed inadequate tools for the complete description of the whole macromolecular structure in that they can be only evaluated for one repeating unit or for a short repeating unit sequence at the best. Moreover, material properties are not only intrinsic to the polymer chemical structure, but they also depend on average characteristics of the polymer, such as, molecular weight, polydispersity index, stereo-regularity, repeating unit distribution. As a consequence, these methods are mostly used for amorphous polymeric materials and are not applied to copolymers, which convey repeating units with different molecular structures. On the other hand, direct treatment of structured data, as it is possible with our RNN model, enables to by-pass the limitations associated with the use of molecular descriptors.

The representation of each polymer was based on the 2D graph of its repeating unit treated as a small molecule. In particular, each repeating unit was decomposed by using the same atomic groups, labels, and priority rules defined for low molecular weight compounds. With respect to the small molecule representation, the most relevant innovation was the positioning of the tree root. Indeed, the root was not placed on the highest-priority chemical group, but on an additional super-source vertex (the group "Start"), not related to the molecular graph. The super-source conveys information on the average macromolecule characteristics through its label. This allows the model to account for both the repeating unit detailed 2D structure and macromolecule average characteristics. In the first application of this approach, we encoded the information of polymer stereoregularity (tacticity) in the super-source label as the fraction of *rr* dyads. It is worth to note that this extension of the representation of molecules to macromolecules is a further point showing the flexibility of a structured representation approach.

The RNN capability of handling structured data together with this original representation of structures results in a good model for predicting polymer properties. The results obtained in predicting the Tg are quite promising. In particular, the RNN model found the Tg-tacticity relationship by treating together polymers with either only one or different tacticity forms. In particular, the potential of the RNN model to take into account the extent and type of stereoregularity of the polymer chains is of paramount importance because of the impact of these features on several properties of the materials. For instance, very often stereoregular polymers are highly crystalline, whereas atactic polymers are amorphous. On the other hand, methods able to correlate the Tg of polymers with their tacticity are lacking in literature.

The results obtained until now highlighted the greater generality and flexibility of our method and of the adopted representation with respect to standard literature methods. In fact, the RNN method can treat small molecules and polymers with the same fundamental approach. In the latter case, the method allows for taking into account also the mean macromolecular characteristics and for the simultaneous handling of polymers for which one or more values of the considered average property exist. Moreover the molecular representation can be naturally extended to the treatment of all types (random, alternating, block) of copolymers.

In closing the overview of the results and of the potential of the proposed method, it must also be mentioned that the analysis of training set outliers was useful to pick out of the data sets the least reliable data of both $\Delta_{solv}G°$ and Tg. For instance, these results have been confirmed by recent literature on polyacrylics and polymethacrylics. This enabled to exploit the RNN, beside prediction, even for data cleaning and data assessment purposes.

CONCLUDING REMARKS

The various examples of application to the cheminformatics domain show the potentiality of a widespread application of structure domain learning methods for the drug delivery and the tissue engineering fields, and, more in general, for biomedical, biochemistry and medicinal chemistry problems. In fact, the processing of structured data by recursive neural networks have been shown to be effective in real-world applications concerning chemical tasks, i.e. the prediction of chemical properties directly from molecular structures. This approach can be seen as a paradigmatic instance of the wider problem of processing structured data by machine learning tools in medical and biological fields. Indeed, biological and biochemical problems are often characterized by quite complex domains where managing of relationships and structures, in form of sequences, trees, and graphs, is important to achieve suitable modeling of the solutions.

Main potential developments concern hard tasks in toxicology, genomics, proteomics, and bioinformatics in general, whenever is natural to find useful structured representation of chemical/biological data or there is the need to capture relevant information such as topological or functional description of the data. Moreover, the flexibility of the structured data learning approach can be exploited to integrate different kind of data arising in medical domain, including genetic, biological, clinical and chemical data. This research can lead to novel approach to *pharmacogenetics* for personalized medicine purposes.

The final aim of the long-term research in such fields is the development of predictive models able to accelerate the discovery of new drugs and new biomaterials, including the studies on their potential genotoxicity, carcinogenicity, or other pharmaceutical toxicity, to anticipate adverse health effect. It is worth noting that the machine learning methods able to deal with structured and complex data domains offers the opportunity to treat in the same compact computational frame such heterogeneous data and problems.

Acknowledgement

The financial support by MIUR Cofin2003 Project "*Integrated multidisciplinary design and realization of biologically active systems for biomedical applications*" is gratefully acknowledged.

REFERENCES

1. Borchardt R.T., Smith P.L., and Wilson G., eds., 1996, "Models for assessing drug adsorption and metabolism", Plenum Press, New York

2. Lee V.H.L., ed., 1990, "Peptide and Protein Drug Delivery", Marcel Dekker, New York

3. Solaro R., Chiellini F., Signori F., Fiumi C., Bizzarri R., and Chiellini E., 2003, J. Mat. Sci. Mat. Med., 14, 705

4. Allemann, E.; Gurny, R.; Doelker, E. 1993, Eur. J. Pharm. Biopharm., 39, 173–191.

5. Kreuter, J. 1994, Nanoparticles, in Colloidal Drug Delivery Systems, Kreuter, J., Ed., Marcel Dekker, Inc., New York, NY, p. 219–342

6. Ringsdorf, H., 1975, J. Polym. Sci., 51, 135

7. Szycher M., Qiu J., and Tanaka M, 2004, in "Kirk-Othmer Encyclopedia of Chemical Technology", Wiley-Interscience, New York

8. R. Shalak, and C.F. Fox, 1988, "Tissue engineering", Alan R. Liss, New York

9. Nerem R.M., 1992, Med. Biol. Eng. Comp., 30, CE-8-CE12

10. Castner D.G., Ratner B.D., and Hoffman A.S., 1990, J. Biomater. Sci. Polymer Edn., 1, 191

11. Tyler B.J., Ratner B.D., Castner D.G., and Briggs B.D., 1992, J. Biomed. Mat. Res., 26, 273

12. Ratner B.D., 1987, in "Treatise on clean surface technology", Plenum Press, New York, p. 247

13. Ziats N.P., Miller K.M., and Anderson J.M., 1988, Biomaterials, 9, 5

14. Bakker D., van Blitterswijk C.A., Hesseling S.C., Grote J.J., and Daems W.T., 1988, Biomaterials, 9, 14

15. Williams D.F., 1987, J. Mater. Sci., 22, 3421

16. Saltzman W.M., 1986, in "Interaction of the blood with natural and artificial surfaces", Dekker Inc., New York, p. 39

17. Baszkin A., 1986, in "Blood compatible materials and their testing", Martinus Nijhoff Publishers, Dortrecht, p. 39

18. Kottke-Marchant K., Anderson J.M., Unemura Y., and Marchant R.E., 1989, Biomaterials, 10, 147

19. Micheli, A., Sperduti, A., Starita, A., Bianucci, A. M., 2001, J. Chem. Inf. Comput. Sci., 41, 202

20. Micheli, A., Sperduti, A., Starita, A., Bianucci, A. M., 2003, In “Soft Computing Approaches in Chemistry”, Sztandera, L. M., Cartwright, H. M. Eds., Springer-Verlag: Heidelberg, p. 265-296

21. Micheli A., 2003, “Recursive Processing of Structured Domains in Machine Learning”, PhD thesis. TD-13/03, Dept. of Computer Science, University of Pisa

22. Sperduti, A., Starita, A., 1997, IEEE Transactions on Neural Networks, 8, 714

23. Bianucci, A. M., Micheli, A.; Sperduti, A.; Starita, A., 2000, Appl. Int. J., 12, 117

24. Bernazzani L., Duce C., Micheli A., Mollica V., Sperduti A., Starita A., Tiné MR., 2004, TR-04-16. Dept. of Computer Science, University of Pisa

25. Duce C., 2005, “Physical chemical methods in the rational design of new materials: QSAR and calorimetric approaches”, PhD Thesis, Dept. of Chemistry and Industrial Chemistry, University of Pisa

26. Kollman, P. A., 1996, Acc. Chem. Res., 29, 461-469

27. Eisenberg, D., and McLachlan, A. D., 1986, Nature (London), 319, 199-203

28. Jorgenson, W. L., and Tirado-Rives, J., 1988, J. Am. Chem. Soc., 110, 1657

29. Jorgensen, W. L., and Tirado-Rives, J., 1995, Persp. Drug. Discuss. Des., 3, 123-138

30. Jorgenson, W. L., and Nguyen, T. B., 1993, J. Comput. Chem., 14, 195

31. Leo, A., Hansch, C., and Elkins, D., 1971, Chem. Rev., 71, 525

32. Smith, R. N., Hansch, C., and Ames, M. M., 1975, J. Pharm. Sci., 64, 599

33. Roberts, M. S., Pugh, W. J., Hadgraft, J., and Watkinson, A. C., 1995, Int. J. Pharm., 126, 219

34. Gratton, J. A., Abraham, M. H., Bradbury, M. W., and Chadha, H. S., 1997, J. Pharm. Pharmacol., 49, 1211

35. Bicerano, J., 2002, “Prediction of polymer properties”, 3rd ed., Marcel Dekker, New York

36. Wiff, D. R., Altieri, M. S., and Goldfarb, I. J., 1985, J. Polym. Sci: Polym. Phys. Ed., 23, 1165-1176

37. Van Krevelen, D.W., 1976, “Properties of Polymers-Their Estimation and Correlation with Chemical Structure”, 2nd ed., Elsevier, New York

38. Koehler, M. G., and Hopfinger, A. J., 1989, Polymer, 30, 116-126

39. Katritzky, A.R., Sild, S., Lobanov, V.S., and Karelson, M., 1998, J. Chem. Inf. Comput. Sci., 38, 300-304

40. Camelio, P., Cypcar, C. C., Lazzeri, V., Waegel, B., 1997, J. Polym. Sci.: Part A: Polym. Chem., 35, 2579-2590

41. Joyce, S.J., Osguthorpe, D.J., Padgett, J.A., and Price, G.J., 1995, J. Chem. Soc., Faraday Trans., 91, 2491-2496

42. Mattioni, B.E., and Jurs, P.C., 2002, J. Chem. Inf. Comput. Sci., 42, 232-240

43. Sumpter, B. G., and Noid, D. W., 1996, J. Thermal Anal., 46, 833-851